東　　海　　大　　學

工業工程與經營資訊研究所

碩士論文

應用資料修補技術於醫療環境之研究

研 究 生：黃啟信
指導教授：王偉華 教授

中 華 民 國 九 十 二 年 六 月

A study of data repairing technique in medical environment

By
Chi-Hsin Huang

Advisor: Prof. Wei-Hua Wang

A Thesis
Submitted to the Institute of Industrial Engineering and
Enterprise Information at Tunghai University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in
Industrial Engineering and Enterprise Information

June 2003
Taichung , Taiwan , Republic of China

應用資料修補機制技術於醫療環境之研究

學生：黃啟信　　指導教授：王偉華 教授

東海大學工業工程與經營資訊研究所

摘要

本研究以更有效的利用原始資料之方法-資料修補機制來探討在醫療環境中可能發生的資料缺失問題，在此提出了一套資料修補機制。首先，應用原本存在於資料中隱含的資訊-特徵項目間之關連性，並且結合了專家知識，使其建立特徵項目之修補架構。再者以類神經網路當中效果良好的一項工具-函式模擬，用以達到補上缺失特徵項目之值。最後應用一種在醫療環境當中常見的現象-少量資料、特徵項目多資料當中，分類效果較為優異的理論方法-支援向量機器，來判斷檢驗本研究之修補機制的結果。

本研究以兩種類型之實例試驗來評估資料修補機制的可行性。第一類型為本身資料結構為完整的資料集型態，目的是方便檢驗資料修補機之成效。以人為的方式使其資料集內容由完整而成為有缺失的資料欄位。並且為了模擬資料缺失的狀態，分別模擬了高度資料缺失、中度資料缺失、以及少量資料缺失三種情境，希望更能切合資料缺失的現實狀態。第兩類型資料便是實際與醫院合作收集之醫療資料，用來做為資料修補機制之佐證。

在本研究進行的實例驗證當中，人為所製造之缺失資料集用資料修補機制來修補其缺失資料空格。以缺失資料跟修補過後之資料用以支援向量機器來做為分類測試，其結果皆為顯著。說明了應用本機制後，對於資料的正確判斷有明顯的增加。

關鍵字詞：資料修補、函式模擬、類神經網路、交互資訊、支援向量機器

A study of data repairing technique in medical environment

Student: Chi-Hsin Huang Advisor: Prof. Wei-Hua Wangk

Department of Industrial Engineering and Enterprise Information
Tunghai University

ABSTRACT

The incomplete-data is the critical problem in the medical environment. In this paper, we provide the data-repairing technique in order to get more information from the primitive data. First, the application exists originally in data in implicit information-relation of the features, and combined expert's domain knowledge, making it create the features to repair the structure. Using the method that Function Approximation is the application in the foundation of Neural Network to fill block of incomplete-dataset up. And the one of excellent method in the classification - support vector machine, judges examination this research it repairs the mechanism of result. This research evaluates the feasibility that data-repairing mechanism is with two of the category types. The first type is oneself the type of data set that data structure is complete; the purpose is a convenience examination data to repair the result of mechanism. Make by factitiousness that it’s the content of data set and become the data field of having the missing. For simulating the status of missing-data, simulated the mass incomplete-data, medium degree incomplete -data respectively, and a little incomplete-data. There are three kinds of scenarios, hoping can also suit realistic status of the incomplete-dataset. The second type data would be to in concert with the medical data of the collections with the hospital physically, using to be used as the data to be the substantial evidence of data-repairing mechanism.

Identification that this research, the incomplete dataset that factitiousness makes repairs the mechanism to fill its blank space of incomplete-data with the data-repairing mechanism. Repair with the incomplete-data later on its data supports the vector machine in order to be used as the classification, the result is obvious. After explaining application this mechanism, have the obvious increment for the right judgment of the data.

Keywords: Repairing-Data, Function Approximation, Neural Network, Mutual Information, Support Vector Machine

誌謝

東海六年的生活一下子就過了，從大學四年及研究所兩年的訓練，經歷了許許多多的故事，令人難忘。

首先，要感謝大學時代的啟蒙老師、研究所的指導教授王偉華老師耐心的指導，無論課業或做人處事態度，總是不厭其煩的提醒與叮嚀，讓我能夠持續地進步往前。

特別感謝丁兆平老師、方曉嵐老師、黃欽印老師及高嘉鴻醫師於口試期間所給予的建議與指正，使這篇論文能夠更加完備。

感謝研究室所有成員及一群好友們，陪我一同走過那段艱苦的日子，有歡笑也有淚水。

最感謝的還是我的家人及女友，無論在精神上及經濟上為我默默的付出及陪伴，一直支持著我。

在此將喜悅的成果與所有的朋友一同分享！

黃啟信　　謹誌於

東海大學工業工程與經營資訊研究所

智慧型系統研究室 ISRG

中華民國九十二年七月

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機與背景

缺失資料(Missing Data)或是稱為不完整資料(Incomplete Data)，一直以來都是在進行資料分析或資料探勘(Data Mining)時一個常見的問題。以目前在進行資料探勘時，遇到缺失資料，往往都是將有缺失的資料項目加以去除。這種作法在處理資料量大的資料集時，例如在從事顧客關係管理的商業行為或是分析網路上所擷取的資料，資料量動輒上萬筆來說，所去除的資料對整體來說影響不大。

但是面對於醫療產業所收集的資料，有時會有著少量、缺失且具有高維度的資料情形發生。一般的資料探勘方法，效果往往不佳。並且在本身即具有著少量資料的情形當中，如果去除了一些發生缺失情形的資料後，對結果的判定造成的影響顯著，例如用於醫療產業所進行資料分析當中時，每一筆資料都是一位病例，並不能完全的加以去除。這樣一來如何去修補缺失資料，使得判斷的結果更加正確，這就是本篇論文當中期望建構的機制來能夠修補缺失的資料，提升系統判斷正確分類的機會。

1.2 研究問題定義

就資料分析的方法來說，面對缺失資料是一種無可避免的挑戰。由於缺失資料的發生原因有時是無可避免的，可能在當時設計資料特徵項目時，所未考量到的資料特徵項目，由於無法重新收集，故之前收集的資料便會產生缺失情形；亦可能在收集的過程當中發生。但因在前小節提及，在醫療產業當中，並不能完全的將資料去除。這時候所面對的資料，要考量的是整體的涵蓋性質。因為由於缺失而去除的資料，很有可能是在這群資料當中，一個很重要的個案。因此在少量、高維度且缺失的資料當中，能做到的便是觀察隱含在其他資料當中的架構，期望能將缺失資料藉由機制來做到修補的動作。

在此將本研究所面臨到的問題，加以條列說明如下：

1. 待分析資料本身各個特徵項目(feature)，是否有著關連性。如果有著關連性時，要如何找尋出來？
2. 一般的分類方法當中，對於處理缺失資料的方法為何？以及本研究欲建立的資料修補機制是否適用？
3. 待資料修補機制建立後，是否可以幫助原本的分類方法增加其擴充性及功能？

1.3 研究方法與步驟

本研究將以下列四個階段，分別為文獻探討、機制建立、實例驗證及論文撰寫，詳細說明如下：

1. 文獻探討部分

   首先以清楚描述缺失資料的類型及影響。且在少量且高維度的資料下，支援向量機器(Support Vector Machine ; SVM)的適用性。在建構修補機制的想法是參考類神經網路(Artificial Neural Network ; ANN)在函式模擬(Function Approximation ; FA)下的應用。以及選擇主要特徵值時所應用的工具- 交互資訊(Mutual Information ; MI)。

2. 修補缺失資料機制的建立

   首先將缺失資料做前置處理，接下來嘗試以交互資訊(Mutual Information)法則或是專家來決定在資料當中，資料特徵項目的關連性。以函式模擬來做為修補缺失資料的方法，以類神經網路為工具加以實現。最後以支援向量機器來做為分類的工具。

3. 實例驗證

   經過實驗設計，來觀察所建構之修補缺失資料機制，對於分類的正確性是否有著相關性。

4. 論文撰寫

   經過前面幾個步驟，最後運用系統化的方式，將其整理並撰寫成論文。

1.4 論文架構

本篇論文架構共分為五章，如圖 1. 1。第一章為緒論，說明本研究期望建立資料修補機制之研究背景與動機，研究目的及範圍，並概要的說明本研究方法。第二章為探討文獻，針對於缺失資料、函式模擬、類神經網路、支援向量機器進行文獻探討回顧。第三章則為研究方法，說明本研究的研究方法及架構。第四章為實證研究，將以實例來驗證本論文設計出來的資料修補機制對於分類的關連性。第五章為結論與未來發展方向。

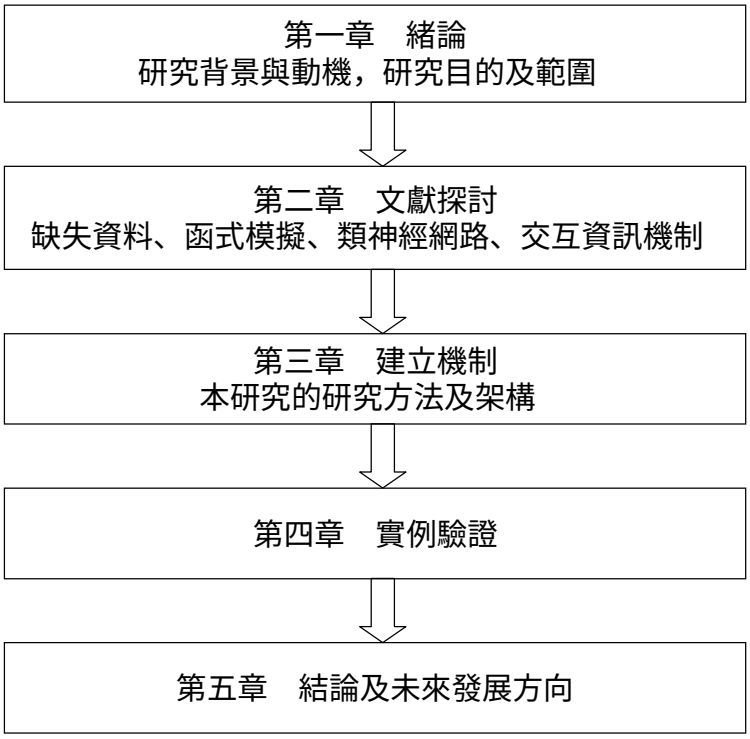


圖 1. 1 本研究架構

第二章 文獻探討

本章節將以五個主題來分別敘述資料修補技術所使用到的研究文獻。首先以不完整資料型態、定義及目前對於不完整資料的修補機制分析討論。接著為了要探討資料欄位間的關連性，本研究以交互資訊法則為主軸加以討論，預期藉由交互資訊法則(Mutual Information)。發現特徵項目間的關連性。接著探討類神經網路(ANN)中函式模擬(Function Approximation)之特性與功能，做為修補資料的機制。最後探討在高維度少量資料表現不錯的分類方法-支援向量機器(Support Vector Machine)來做為分類的工具之說明。

2.1 不完整資料（Incomplete Data）

2.1.1 不完整資料之簡介

不完整資料（Incomplete Data）在資料分析或是資料探勘當中，往往是一件難以去處理的課題。因為不完整資料在應用層面上是經常面臨到的資料型態，而且在理論基礎上亦面臨許多難以架構的困難點。在醫療環境當中，往往發現收集到的資料所面臨的就是資料量不足及缺失的問題。資料不足與缺失是一體兩面的事，因為缺失的資料不能使用時，便會導致資料量不足的現象產生。

2.1.2 不完整資料的分類

不完整資料出現的情況可劃分成相當多種類，而每一種不完整資料的處理方式與技術分析亦有所不同。將其分類做一簡要說明(陳淑婷，民 91)。

(1) 遺失值（Missing Values）：

在資料處理過程中，因實驗單元之遺失，造成訊息無法完全取得。此種情況通常是出現在生物、醫學或科學工程實驗的資料上，因為試驗單元（experimental unit）在試驗過程中因為某些原因而退出試

驗，如死亡、遷移…等因素。或是在序列的資料測試中，失去了某幾次的測試資料（包括有意或無意），如設備出現問題，造成無法完整取得規劃中的資料，所以遺失值（Missing Values）所表達出的情況為「某些觀察向量的遺失，或是某些向量中只取得其中一部分的資料」。

(2)無回應資料（Non - Response Data）：

此種資料的殘缺結構和遺失值（Missing Values）是相同的。但 Non - Response Data 是特定指在調查時的資料沒回應之情形，如做問卷調查時，發出了一千份問卷但是卻只回收三百份，而未回收的七百份資料則稱為 Non - Response Data，若在回收的問卷中並非所有問卷上的問題皆有被回答，則此種資料則稱為 Item Non - Response Data。如果未回應的資料佔了相當大的比例，則會對此項調查的品質造成傷害。因為無回應的資料（Non - Response Data）本身可能存在某種特殊結構，如此一來會使得回收資料所提供的訊息產生偏誤，這是所謂的目標母體與抽樣母體之間的差異。

(3)設限資料（Censored Data）：

此種不完整資料是指當某個資料出現時，只知道這個資料落在的「某個範圍內」或「某個區間內」，但其實際數值為何，則因為無足夠能力、時間去作判斷而無法得知。例如我們要測試一百個燈泡的壽命，首先將全部同時開啟，在經過一千個小時之後已知有四十個燈泡已燒毀，將其壽命記錄下來。然而另外六十個燈泡，只知道其壽命都大於一千個小時，但是無法確切得知其壽命為何，在此情形下，稱這六十個觀察資料為設限資料（Censored Data）。

在醫學上的測試，常會聽到某種疾病在接受某種醫學治療後，其在十年之內未再復發，此種資料亦可稱為設限資料，因為十年後此疾病是否會再復發無法得知。而如果已接受此醫學治療的病人，其五年後死於其他原因(如意外)，則對某種病的復發事件而言仍屬於設限資料，只因無法再對病人進行觀察。

(4)截切資料（Truncated Data）：

截切資料與設限資料主要的差別為訊息可能存在或出現過，但卻無法得知，亦即，在一個母體中有部分樣本完全不可能被觀察到，例如人類的耳朵並不能聽到所有的聲音，因為當聲音頻率超過某個範圍時耳朵就無法聽見。所以這個聲音訊息即使曾經出現過，人們也完全不知，而此種資料型態就稱為截切資料（Truncated Data）。

(5)隱藏式變數（Latent Variable）

隱藏式變數亦稱為不可觀測之變數，通常是指在模型化的過程中「必須」存在的變數，但是這些變數只是觀念上的存在，事實上，無法觀察到它的數值反應為何。而在因素分析中稱為此種變數亦稱為共同因子（Common factor），而這種共同因子是無法直接由觀察得知。

2.1.3 目前對於缺失資料的處理方式

目前有三種的缺失值處理方式，分列如下(Ng *et al.* , 1998)

(1) 刪除有缺失值的整筆記錄(橫向)

最簡單和最直接的方法是在整個資料分析的過程中，只要包括有完整資料的記錄(record)。

(2) 忽略含有缺失值的欄位(縱向)

第二種方式是將包含有缺失資料的特徵項目(feature) 欄位在資料分析的過程中直接忽略。

(3) 第三種補足缺失資料的方式是採用不同的技巧方法將缺失資料補足。

以下介紹常用到的取代缺失值的方式：

a. 記錄置換法(Record substitution)：

這種方式是利用其他資料集(data set)中擁有完整資料的記錄來取代目前擁有缺失資料的記錄。

b. 平均值置換法(Mean substitution)：

平均值置換的方式是將同一特徵項目欄位中的平均值計算出，再以此平均值填入原有的缺失值中。

c. 徹底裝飾法(Cold deck imputation)：

從過去的資料集中，同一特徵項目欄位資料中學習出一個固定不變的常數值，此後的缺失值出現在同一特徵項目欄位中時，再利用此常數值填入。

d. 迴歸法(Regression imputation)：

迴歸分析法通常被用來預測缺失值，迴歸分析法是以資料集中其他特徵項目為基礎藉其特徵項目間的關係或關連性，用來預測出其他欄位特徵項目的值。

本研究採用的方法與上述四項方法略有差異，2.2 節將會來探討尋找特徵項目間關係的方法。

2.2 熵理論與共同資訊

早期資訊理論主要用以解決資料傳輸上的資訊計量分析工作，資料從資訊源到目的地的資料傳輸過程中，每一步驟均可能受到不可控制的因素影響，使資料的傳輸過程發生錯誤。

為維護資料及分析工作必須建立適當的指標，以衡量資料傳輸的資訊量多寡。一般常用的資訊衡量指標有三種 (Blahut, 1987)：

(1)熵 (entropy)

(2)共同資訊 (mutual information)

(3)識別力 (discrimination)

將僅敘述熵、共同資訊這二項在本研究有相關的指標。

若隨機變數 X 對系統產出有影響力，則將 x 視為一個訊息通道 (message channel)。若變數 x 的值域可等分成 m 個區間，則任一區間 x_k 發生的機率為 p_k。當 x= x_k)發生，熵值之定義如下：

$$I(x_k) = \log(\frac{1}{p_k}) = -\log p_k \qquad \text{(式 2. 1)}$$

事件 x= x_k 尚未發生前，I(x_k)可解釋為系統輸出的不確定性 (uncertainty)；當事件發生後，系統輸出的不確定性已消失，則 I(x_k)

解釋為該事件發生後的資訊獲得 (information gain)。I(x_k)具有下列物理特性 (Haykin, 1999)：

1. I(x_k) =0, for p_k =1
   若某一事件 x_k 的發生機率 p_k 為 1，則無任何不確定性可言，因此無法從該事件的發生獲得任何資訊。
2. I(x_k) ≥0, for 0≤ p_k ≤1
   若某一事件 x_k 發生，不是獲得資訊 (>0)，就是沒有獲得任何資訊 (=0)，但絕不會造成資訊損失。
3. I(x_k) > I(x_j), for p_k < p_j
   某一事件發生的機率愈低，一但該事件發生，則觀察者所獲得的資訊愈多。若此事件發生，其機率越低則獲得資訊越多。

I(x_k)的大小視機率值 p_k 而定，隨機變數 X 的熵值之期望值 H(X) (式 2.1)。若有兩隨機變數 X 與 Y，則系統的條件熵值之期望值為 H(X|Y) (式 2.2)。H(X|Y)表示觀察到變數 Y 值後，系統對 X 輸出所剩餘的不確定性。H(X)-H(X|Y)為透過觀察系統輸出 Y 值後，所降低的不確定性。I(X,Y)為二變數 X 與 Y 之間的共同資訊計算方式。

$$H(X) = E[(I(x_k)] = \sum_k p_k I(x_k) = -\sum_k p_k \log p_k \qquad \text{(式 2. 2)}$$

$$H(X|Y) = H(X,Y) - H(Y) \qquad \text{(式 2. 3)}$$

$$I(X,Y) = H(X) - H(X|Y) \qquad \text{(式 2. 4)}$$

當二隨機變數 X 與 Y 的相關性愈大，二者的共同資訊量 I(X;Y)愈高，透過其中一者的觀察可降低另一變數輸出值的不確定性。I(X;Y)具有下列物理特性 (Cover and Thomas, 1991; Gary, 1990)：

X 與 Y 之共同資訊具有對稱性 (symmetric)，即 I(X,Y)=I(Y,X)

X 與 Y 之共同資訊不為負值 (nonnegative)，即 I(X ,Y) ≥0

X 與 Y 之共同資訊可以 Y 之熵值表成下式 2.5

$$I(X,Y) = H(Y) - H(X \mid Y) \qquad \text{(式 2. 5)}$$

類神經網路將輸入向量轉換成輸出向量的過程，可視為一個訊息轉換的過程。將每一特徵項目視為訊息輸入通道，而類神經網路的判定結果為系統輸出通道。因此，上述理論 (熵理論與共同資訊)可應用於評估特徵項目之間的關連性。

2.3 類神經網路-基本概念

類神經網路(Artificial Neural Network)顧名思義可以說是一種類似、模仿人類實際神經運作的過程，其由許多個別的神經元(Neuron)構成一個網路，並透過這個網路來處理並解決問題，以下就類神經的基本概念及本研究所要採用的其中一個網路模式 - 倒傳遞網路模式(Back propagation)以介紹。

2.3.1 神經網路之組成元素

人類的神經系統是由神經元(Neuron)所組成，它們的數量非常龐大，約有個10^{11}神經元互相連結成約10^{15}條傳輸的路徑，而每個神經元能夠接收、處理並經由神經的路徑傳輸電化學信號，進而形成一個腦神經的通信網路系統，而一個神經元的組成元素則又可細分如下。

a.神經細胞核（Soma）：

為細胞神經的中心體，它的主要作用至今仍尚未被徹底的瞭解，粗略說明是將神經樹所收集到的信號在此做一加總後透過轉換的過程，再經由神經軸將信號傳送至其他神經細胞中。

b.神經軸（Axon）：

連接在神經細胞核上，用來傳送由神經細胞核產生的信號至其他的神經細胞中。

c.神經樹（Dendrites）：可分為輸入神經樹與輸出神經樹兩種，主要也是用來傳送信號至其他神經細胞。

d.神經節（Synapse）：

神經節為輸入神經樹與輸出神經樹相連接的點，它表示兩個神經細胞間的聯結強度，在人工神經元上將此聯結強度以一數值來表示，稱之為加權值。圖 2. 1 為生物神經元圖示說明。

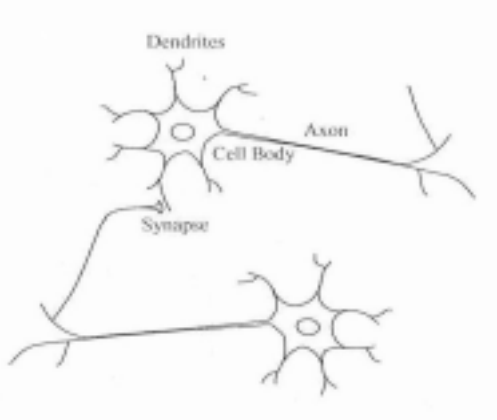


圖 2. 1 生物神經元

2.3.2 類神經網路之形成概念

由於人類神經網路的運作情形是接收某一刺激或訊號之後，藉由神經樹、神經軸傳輸到神經核進行處理之後，假設此訊號或刺激夠強則再傳輸新的脈波訊號給其他的神經細胞來做為反應動作的依據，所以在模仿此運作過程時也必須考量各個層節的特性，而值得注意的是當訊號透過神經節後，會由於神經節的加權值關係而使的訊號大小產生變化，圖 2.2 為一個人工神經元的模型，其具有多個輸入值以及一個輸出值，而輸入與輸出之間一般都以下(式 2. 6)加以表現之。

$$y(t) = f\left(\sum_{i=1}^{n} w_i \cdot x(t) - \theta\right) \qquad \text{(式 2. 6)}$$

w_i：表示模仿生物神經細胞的神經節加權值。

x_i：表示接收之輸入訊號。

θ：表示模仿生物神經細胞的細胞核門檻值（偏權值），即輸入訊號的加權乘積且必須大於此門檻值，才會有訊號被傳至其他人工神經元中。

t：表示時間。

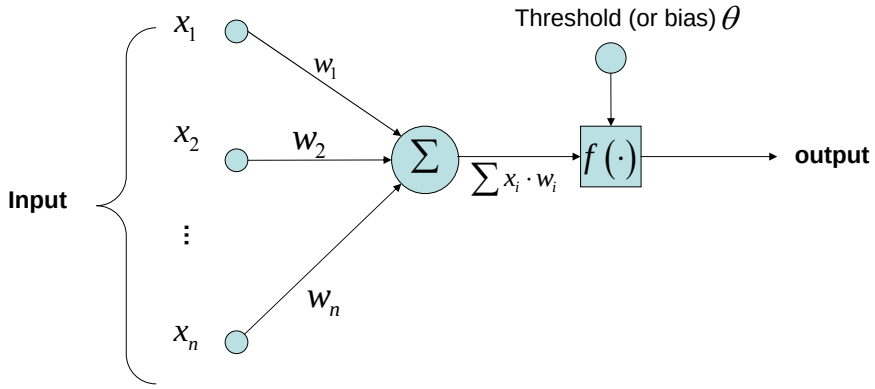


圖 2. 2 人工神經元模型

2.3.3 類神經網路之類型

一般來說類神經網路應用非常廣泛，舉凡工程、醫學、生技、管理、機械等方面都有相關的研究及應用，在這幾個領域之中的應用又以分類（Classification）、預測（Prediction）、最佳化（Optimization）方面的應用為主流，不同目的及不同資料型態就會有不同的網路型態以供應用，大致上可以將它們分成四個種類，監督式學習網路（Supervised Learning Network）、非監督式學習網路（Unsupervised Learning Network）、聯想式學習網路（Associate Learning Network）、最適化應用網路（Optimization Application Network）等。

2.3.4 類神經網路架構

類神經網路的基本架構大體可分成兩大類，一為迴歸型網路（Recurrent Network），另一為前授型網路（Feed-forward Network）；其分述如下：

（1）迴歸型網路：

在此網路型態中人工神經元彼此相連。對每個神經元而言它的輸出連接至所有其他的神經元，而它的輸入則來自於所有其他神經元的輸出，換言之網路中的每一個神經元都是平行的接受所有的神經元輸入，再平行的將結果輸出到網路中其他神經元上，這類型的網路以霍普菲爾網路（Hopfield Neural Network）為代表。

（2）前授型網路：

此種網路是一種階層式網路，是由許多的層（ Layer）所組成，如輸入層（ Input Layer）、隱藏層（Hidden Layer）、輸出層（Output Layer），每一層中都包含一個以上的神經元，而不同層間的神經元則互相聯結，在信號傳輸方面則是單向式，由輸入層至隱藏層再到輸出層，而這類型的網路以倒傳遞網路（Back propagation Network）最具代表性。

2.3.5 輸入與輸出值

在探索事件發生的原因時，通常都會以應變數與自變數之間所存在關係來解釋之，只是兩者之間所存在的關係在此是以類神經網路模式來加以表達兩者之間的關係。

所以在構建網路模式時的第一要件便是確立出目標，而目標就是輸出值，再來便是找尋解釋變數，此即輸入變數，就像其他傳統上的統計方法一樣，要能有效的解釋輸出值則必須對於輸入變數做一謹慎的選擇以反應其特性。

2.3.6 轉換函數

一般在輸入解釋變數數值之後倒傳遞網路會先將其乘上權重值再做一加總的動作，此時再將此加總完的總數與門檻值加以比較，若此數值大於門檻值則會再透過一個函數將所得到的輸入轉換成另一數量的刺激輸出，此函數便是轉換函數(Transfer Function)，轉換函數的設立主要是要模仿神經元在受到刺激之後所產生的反應大小；而此種函數又可分成兩種型態，一為連續形轉換函數如雙曲線正切函數（Hyperbolic Tangent Function）、雙彎曲函數（Sigmoid Function）等等，另一為離散形轉換函數如線性函數（Linear Function）、階梯函數等等一般來說較複雜的事件大多適用連續形的轉換函數比較能符合現實的情況，但若事件本身與解釋變數是具有線性的現象時，則利用傳統上的線性迴歸方法便能有效解決問題而不需要用到如此複雜的網路概念。

2.3.7 學習過程

類神經網路所具有一獨特之特性便是其與人腦一樣具有學習的功能，透過不斷的學習來達到預期的目標，而所謂預期的目標主要是指要使網路訓練出來的數值（Real Output）與實際上期望的數值（Desired Output）能相當接近或是在某一可忍受的範圍之內，亦即使兩者之間的誤差能最小化，要達到此目標便是透過不斷的訓練從誤差之中去學習並回饋到權重值的修正，式 2.7 便是用來衡量此一誤差的公式，稱之為能量函數（Energy Function）。

$$E = \frac{1}{2}[\sum_{i}\left(T_i - O_j\right)^2] \qquad \text{(式 2. 7)}$$

其中，

T_i 表示第 i 期望輸出值

O_i 表示第 i 網路輸出值

2.3.8 學習法則

類神經網路透過學習過程來判定網路是否要繼續學習，或者是已達到可容忍的誤差範圍而停止學習，而倘若網路未達到標準而需繼續學習以調整權重值時，該如何去調整、調整多少則有賴於學習法則的設定，一般來說學習法則就是用來調整權重值以使網路所得的輸出值與期望的輸出值之間差距縮小，大部份都是應用最陡坡降法的概念來從事權重更新的動作，而大部份在做權重更新都需要三個元素來做計算以推導出權重移動的方向以及大小，此三元素分別為原始權重值、原始輸入值、原始輸出值，之後再根據更新後的權重繼續做下一步的更新，如此重複直至滿足能量函數在某一容忍範圍之內才停止。

2.3.9 回想過程

在此一過程之中，網路接受外來的輸入，並依之前所獲得的模式回想演算法，在經反覆運算之後由輸出層神經元將結果輸出。

2.3.10 類神經網路-倒傳遞網路

類神經網路之運作方式為事先設計或由隨機產生的網路權重初始值。將資料輸入網路後，藉由網路能量函數 (energy function)、目標函數 (objective fuction)或績效量測 (performance measure)的變化來調整權重值。大部分網路學習為調整網路權數，藉由能量函數作為調整或設計權數之依據。當能量函數值愈大時，表示網路運作的實際輸出值 (Oi)與原系統運作輸出值 (Ti)之差異愈大，即網路無法表現真實系統之行為。反之，能量函數愈小，則網路詮釋原系統之行為能力愈強。一般類神經網路採用驟降式搜尋法 (steepest descent method)逐步尋找網路的區域能量函數極小值 (Tarun, 1990)。

類神經演算法則中，最常運用的是倒傳遞演算法，倒傳遞演算法可分為兩個階段 (Dayhoff, 1990)：

1. 前饋階段 (forword-propagation step)：由輸入向量於網路之輸入層，繼而前向傳遞至系統的隱藏層。在前向階段最主要的工作為根據網路內部參數，計算網路的輸出值。

$$E = \frac{1}{2}[\sum_{i}\left(T_i - O_j\right)^2] \qquad \text{(式 2.8)}$$

2. 後饋階段 (backward-propagation step)：通常延續前饋階段後進行。首先，計算網路的實際輸出值與期望輸出值的差距量 δ。根據此差距量修正網路輸出層與相連結之隱藏層間的權重值，此修正程序往後延續至隱藏層與相連結的輸入層。此階段主要是修正網路內部參數，以降低網路在前向階段所產生的輸出誤差。

類神經網路的網路結構中，每一層的隱層元個數並無定論。若隱藏元個數太少，則無法有效分類；若隱藏個數太多，又會造成網路運

算與記憶體之負擔，也有可能降低網路的學習效果。根據研究顯示，若隱藏元個數足夠，則只需要單階隱藏層就可處理大部分的函數模擬問題 (Nielsen, 1989; Stinchombe and White, 1989, Hornic et al., 1989)。

2.4 支援向量機器(Support Vector Machine)

支援向量機器(Support Vector Machine)(Vapnik 1998)是由 Vladimir Vapnik 在 1979 年開始研究發展於 1998 發表的一種分類方法。為一種機器學習(machine learning)的方法，可應用於分類(pattern classification)以及迴歸(regression)。它可以同時降低訓練錯誤(training error)以及測試錯誤(testing error)，因此是最近機器學習領域很熱門的一種方法。

2.4.1 SVM 基本表示法

一般對於 SVM 當中因子基本表示法如下 ：

x：是一個向量，描述一筆資料的各個特徵值(attribute)

y：為 +1 或 -1，表示兩種類別(class)

D：為決定函數(decision function)

$$D(x)=(w\cdot x)+w_0 \qquad \text{(式 2. 9)}$$

且 $w=\sum_i y_i \cdot x_i$

2.4.2 圖形分類(pattern classification)

首先要給定一筆資料 x，判定該資料是屬於哪一個類別(以兩類為例，+1 或-1)。在給定的訓練資料(training data)中，找出一個超平面(hyperplane)，能夠把資料區隔開來，如下圖 2. 3 所示：

找到一個超平面，也就是找到相對應的 w 和 w_0，將訓練資料分開來，以給定一筆測試資料(test data) x，根據決定函數。若 $D(x)>0$，則將該筆資料歸類為+1，若 $D(x)<0$，則將該筆資料歸類為-1。

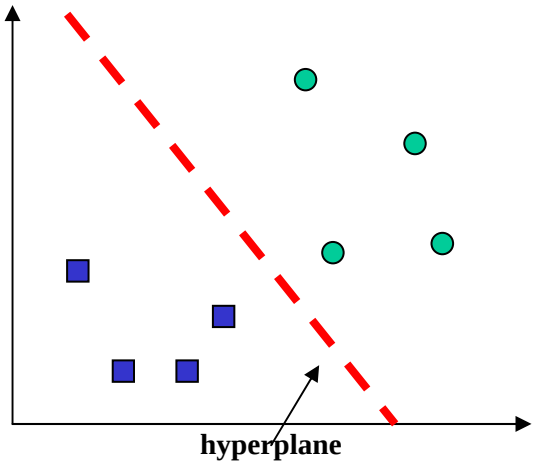


圖 2. 3 SVM 說明圖 1

2.4.3 邊界(margin)及 Support Vector

Hyperplane 的 D(x) =0，而距離 hyperplane 最近的資料點就是所謂的 support vector，因此將 support vector 代入 Decision function 的值為 D(x) = 1 與 D(x) = -1，也就是圖 2. 4 中的二條虛線。而其餘資料點代入 Decision function 必定大於 1 或小於-1，若資料點代入 Decision function 大於 1 為其中一類，小於-1 則為另一類。

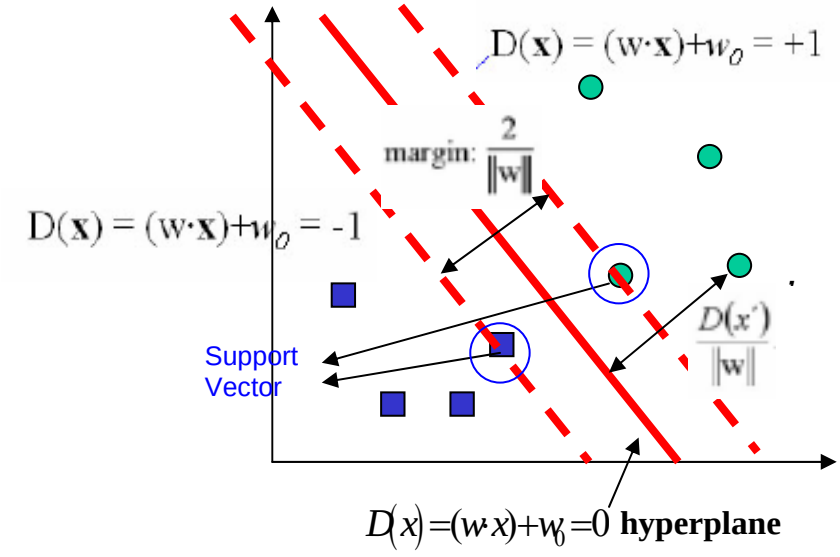


圖 2. 4 SVM 說明圖 2

Support vector 與 hyperplane 的距離為 1/||w||，由於經由計算可以得到可能不只一個 hyperplane 來將這些資料分開，因此 SVM 會找到

margin = 2/||w|| 為最大的 hyperplane，則此 hyperplane 就是最理想的 hyperplane，也就是說資料可以依最理想的 hyperplane 做最明確的分類。

2.4.4 特徵空間(feature space)學習

在大部分的情況下，資料沒有辦法被線性的分類，因此要把資料映射(map)到特徵空間，見圖 2.5，意味著若資料無法在所在的維度下明確的分類，SVM 會將資料轉換至高維度之後再做分類，轉換公式如下：

$$\Phi : R^n \to R^m, m > n \qquad \text{(式 2. 10)}$$

觀察對偶問題，發現在對偶問題中，資料的處理都是用到向量內積(inner product)，因此若要在特徵空間學習，只要能計算出資料在特徵空間中的內積值就可以了，並不需要直接把資料映射到特徵空間。而將資料轉換至高維度之後，在計算式 2.11 時就必須耗費時間來做內積的運算，因此 SVM 便會定義 kernel function 來簡化內積運算，以加快運算的速度，Kernel function 定義如下：

$$k(x, y) = \Phi(x) \cdot \Phi(x_j) \qquad \text{(式 2. 11)}$$

而 SVM 中所定義的 kernel function 有下列幾種：

Simple dot : $k(x, y) = x \cdot y$ (式 2. 12)

Polynomial : $k(x, y) = (x \cdot y + 1)^p$ (式 2. 13)

Radial basis function : $k(x, y) = \exp\left(\frac{-\|x - y\|^2}{2\delta^2}\right)$ (式 2. 14)

Sigmoid kernel : $k(x, y) = \tanh(k(x \cdot y) - \Theta)$ (式 2. 15)

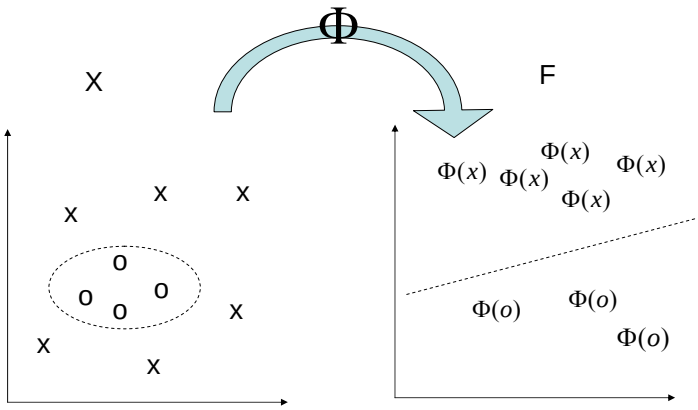


圖 2. 5 空間轉換圖

2.4.5 LIBSVM

LIBSVM(Chih-Jen Lin,2003)全名為 A Library for Support Vector Machines 為林智仁博士所開發的一套 SVM 軟體，本研究將以 LIBSVM 做為分類工具，做為分類結果的依據。

第三章 研究方法

3.1 資料範圍

在本研究將探討缺失資料的修補方法，先在此定義處理資料的類型。

(1) 資料型態將為數值型態，由於本研究的方法不能處理非數值型態的資料，所以要面對的資料集，必須先行轉換為數值型態。舉例來說，如性別特徵值原本為男或是女來辨識，必須轉換為 0 跟 1 來辨別。

(2) 缺失資料集的型態為隨機缺失型態，由於一般來說缺失資料（Missing data）所表達出的情況為「某些觀察向量的遺失，或是某些向量中只取得其中一部分的資料」。

(3) 在此資料集當中，特徵值之間並非完全獨立。在特徵值當中有著關連性或是函數性質存在。並且在這些有著關連性資之特徵值是可藉由專家來加以設定其關連性。

3.2 研究工具

本研究原始資料為文字檔格式、Microsoft Access 和 Microsoft Excel 的資料，在此先以 Microsoft Excel 作為轉換格式的工具，將處理過的資料輸入資料庫中，再匯入 MATLAB 做為各項研究步驟的輸入資料。程式平台架設於 Windows XP Professional，先以 MATLAB 作為整合介面的工具。MATLAB 應用軟體為一套應用於數值計算、數據視覺化及動態模擬的軟體，故本研究程式當中特徵關連選擇、交互資訊(Mutual Information)以及資料表縮減的機制將以 MATLAB 的函式來呈現，建立類神經網路是利用 MATLAB 的類神經網路工具箱(Neural Network Toolbox)，建構與訓練網路特徵值的權重值。支向機的分類部分將以 LIBSVM 來作為工具使用，最後將其結果記錄之。

3.3 研究機制建立

本研究方法將分為兩部分：第一部分為資料修補機制，第二部分為資料分類及驗證機制，本研究的資料流程如圖 3.1。

在第一部分（資料修補機制）當中，將會依照以資料收集、資料前處理、資料特徵選定、拆解資料集、資料修補方法（類神經網路）、修補後資料合併順序加以說明。

在第二部分（資料分類及驗證機制）當中，將討論未經由資料修補機制以及經由資料修補機制的分類情形及方法探討，最後以實例來驗證其效果。

以圖 3. 2 研究機制流程圖來說明。

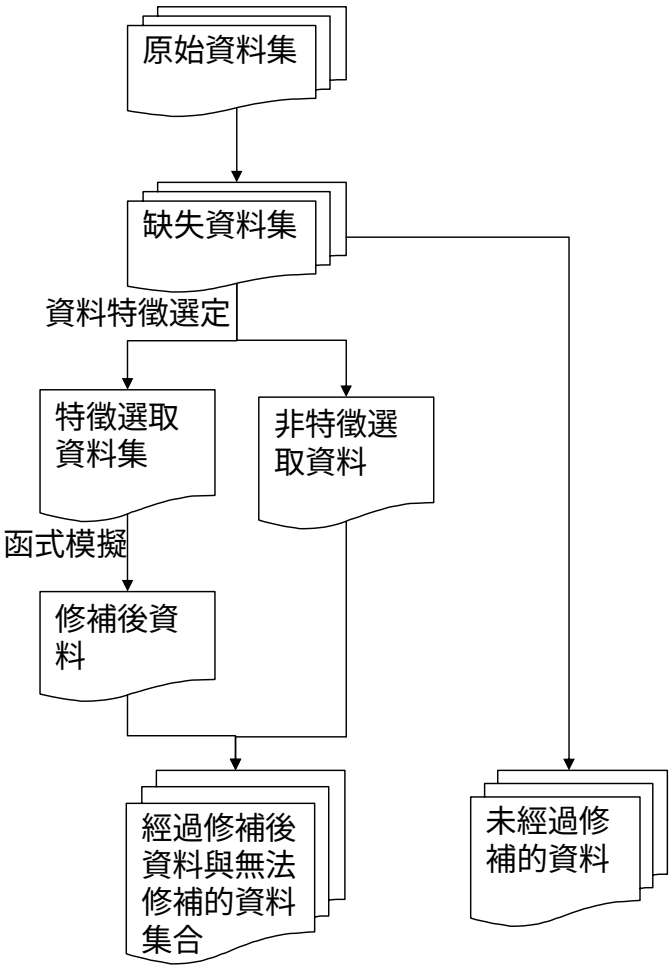


圖 3. 1 研究資料流程

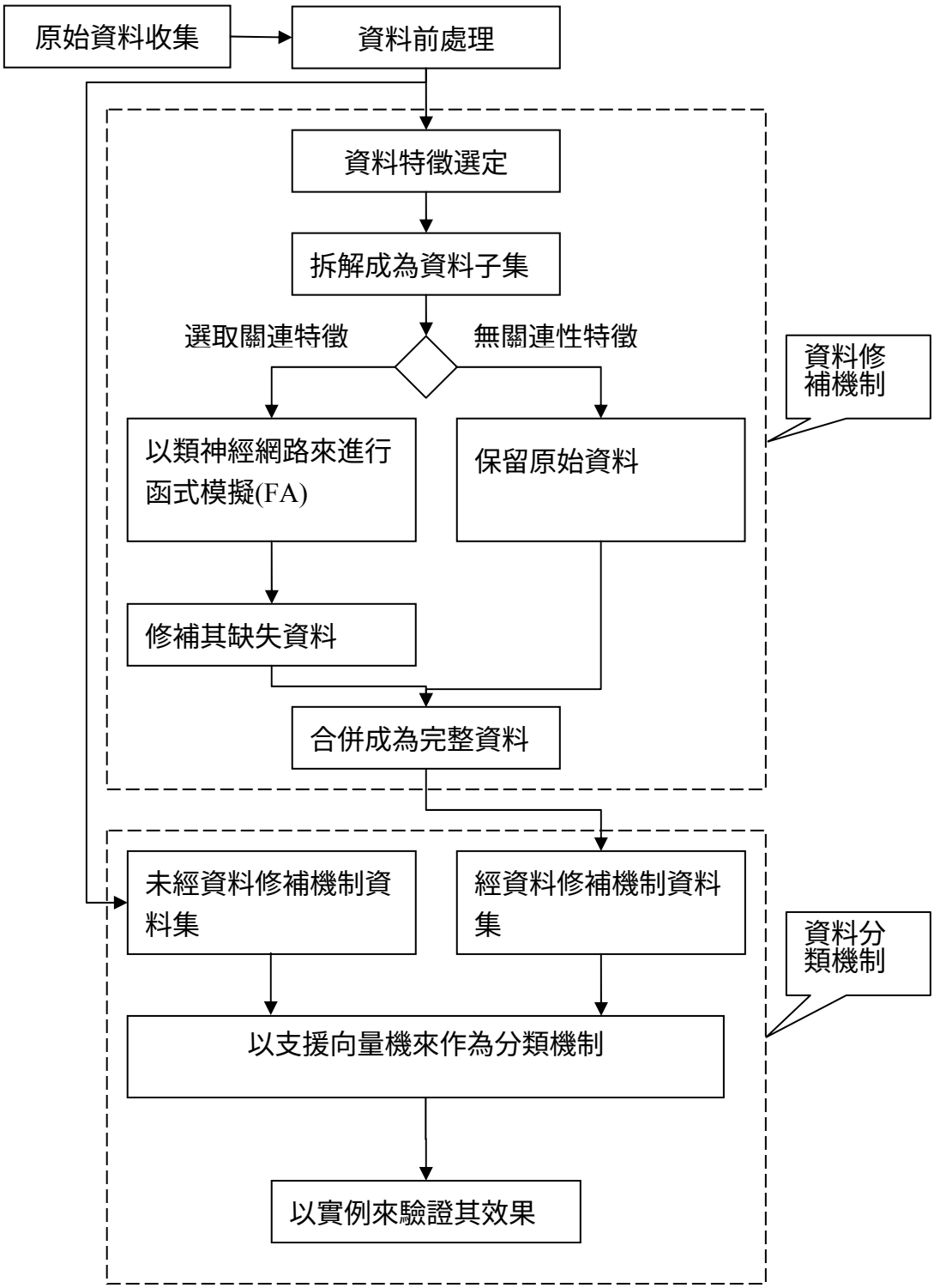

原始資料收集
資料前處理
資料特徵選定
拆解成為資料子集
選取關連特徵
無關連性特徵
以類神經網路來進行函式模擬(FA)
保留原始資料
資料修補機制
修補其缺失資料
合併成為完整資料
未經資料修補機制資料集
經資料修補機制資料集
資料分類機制
以支援向量機來作為分類機制
以實例來驗證其效果

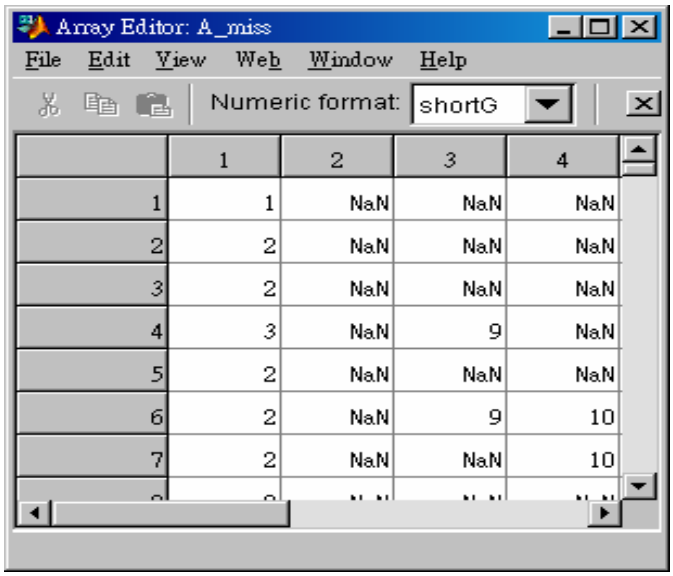
圖 3. 2 研究機制流程圖

3.3.1 資料收集及資料前處理

在資料收集開始，所收集的資料檔案為不同的格式，有文件檔、Microsoft Excel 資料表及 Microsoft Access 資料表的格式。至於資料內容部分，對於特徵項目的描述、對於缺失資料的表示方式、資料內容的本身格式不盡相同。

由於使用工具為 MATLAB，對於 Microsoft Excel 資料表有著已設定好的匯入模式，故在此先加以轉換資料檔案格式為 Microsoft Excel 資料表。內容部分將在這裡以 Microsoft Excel VBA 對原始資料加以整理，將文字或是符號部分加以轉換為數值資料，以方便後續處理程式。

至於資料缺失部分，在 Excel 資料表當中該欄留下空白。在使用 MATLAB 匯入時，資料呈現結果將為 NaN(Not a Number)，如圖 3.3 所示。



	1	2	3	4
1	1	NaN	NaN	NaN
2	2	NaN	NaN	NaN
3	2	NaN	NaN	NaN
4	3	NaN	9	NaN
5	2	NaN	NaN	NaN
6	2	NaN	9	10
7	2	NaN	NaN	10

圖 3. 3 缺失特徵值在 MATLAB 中表示方式

3.3.2 資料特徵選定 - 專家參與

本節將以領域知識專家進行特徵項目篩選，來達到選取資料特徵的目的。首先以舉例來說明資料特徵選定的過程及方法。

缺失型態資料內容(表 3. 1)表示：

表 3. 1 缺失型態資料內容

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8
A	2	9	5		4	12	4	10
B	2		5	5	4			2
C	3	9	5	4	4		4	2
D	2	9	5	4			5	1
E	2		5	5	4			2
F	2	9		4	5		5	2
G	2	9	5	4			5	

註: 灰色部分表示該特徵值為資料缺失。

由特徵值來看，整組資料當中屬於完整(complete)僅有 Feature1，其他特徵值為皆有缺失(incomplete)的情形發生。

如果以一般的作法來論定資料完整性的話，那僅有 Feature1 可以存留下來，並且沒有任何一筆資料是完整的。但是失去了其他特徵項目的資料，那分析出來的，可能不是所想要的結果或是無法正確的找出隱含在資料內的意義。

在本研究中提出一個方法，就是藉由著專家來幫助選定特定特徵項目來做到關連的動作。如專家表示在 Feature1、Feature3、Feature4 這三個特徵項目有著特定的關連性。又 Feature5、Feature7、Feature8 亦有關連性。因此在這裡產生了兩個子資料集(表 3. 2 及表 3. 3)。

表 3. 2 子資料集 1(Feature1,Feature2,Feature4)

	Feature1	Feature3	Feature4
A	2	5	
B	2	5	5
C	3	5	4
D	2	5	4
E	2	5	5
F	2		4
G	2	5	4

在子資料集 1(表 3.2)當中，由於專家經驗告知了這些特徵有著關連性，故匯集成子資料集。與母資料集(表 3. 1)有著兩點不同處。

第一，由於特徵項目變少了，資料的複雜度也隨之降低。第二，缺失的情形比母資料集更為減少。以本例來說明，原本完整的資料數為零，經由專家挑選的方法來處理，完整資料變為 4 筆。

表 3. 3 子資料集 2(Feature5,Feature7,Feature8)

Feature5	Feature7	Feature8
4	4	10
4		2
4	4	2
	5	1
4		2
5	5	2
	5	

在子資料集 2 (表 3.3)當中也與子資料集 1 (表 3.2)相同。特徵項目由 8 個降為 3 個，完整資料由 0 筆至 4 筆。

3.3.3 資料特徵選定 – 交互資訊(Mutual Information)法則

由 3.3.2 小節說明以專家經驗及其專業知識可以得到一種將資料特徵值縮減的方法，優點是不經過任何處理便可以達到選取到有關連的資料特徵值。但是有可能該專家選取之範圍無法將資料當中所隱含的關連性選取出。此時可以使用交互資訊(Mutual Information)法則來計算出各個特徵值資料的關連性。進而選取子資料集加以分析。

根據 2.2 當中所提及的方法。在本研究加以使用。使用步驟如下：

(1) 計算各欄位（特徵值項目）之間 MI 值。

(2) 選定一門檻值。

(3) 若無任一值達到門檻值的條件時，則選取出 MI 較大的固定特徵項目。

(4) 選取出有關連的特徵項目。

在此方法當中，對於門檻值的設定有兩項設定標準。首先設定其門檻值數值。若沒有任何一項項目符合門檻值時，則選取排序過後，MI 數值較大的固定特徵項目。

舉例說明。假設表 3. 4 為一資料集的 MI 對照表，假設門檻值設定為 0.1 時。選取過程如下所示：

(1)在若要選取與 A 有著較高相關性的特徵項目則選取 B 及 C。

(2)在若要選取與 B 有著較高相關性的特徵項目則選取 A。

(3)在若要選取與 C 有著較高相關性的特徵項目則選取 A。

故經過交互資訊法則選取過後。若要修補 A 特徵項目時，A_{ B C}為所選取出之子集合。相同的，若要修補 B 特徵項目時，B_{A}為所選取出之子集合。

在這當中 A_{ B C}所表示的意義為 A 特徵與 B 特徵有著關連性且 A 特徵與 C 特徵有著關連性，B 特徵與 C 特徵不一定有著關連性 。

表 3. 4 MI 範例說明一

	A	B	C
A	-	0.11937	0.1004
B	0.11937	-	0.059642
C	0.1004	0.059642	-

接下來再以一例子說明未到門檻值的情形。表 3. 5 亦為一 MI 參照表，假設門檻值為 0.1。在選取與 A 有著較高相關性的特徵項目時，發現並沒有任何其他特徵項目達到門檻值的要求。因此啟動第二個設定-選取數值較大的固定特徵項目。

故對於 A 特徵項目來說，則選取 A_{D E}兩對特徵項目來做為選取出之子集合。

表 3. 5 MI 範例說明二

	A	B	C	D	E
A	-	0.016558	0.024433	0.0348	1.302
B	0.016558	-	0.036982	0.0379	0.6686
C	0.016558	0.036982	-	0.026796	0.034679
D	0.024433	0.036982	0.026796	-	0.037348
E	0.0348	0.0379	0.034679	0.037348	-

結合 3.3.2 及 3.3.3 以上兩種方法所得之特徵項目集，預期能達到資料面與領域知識之整合。

3.3.4 資料修補機制 – 以類神經網路用來進行函式模擬

首先由 3.3.2 以及 3.3.3 這兩節可以得到，由原始資料集所分解出來的子資料表，在此由於方便說明便在此將表 3. 3 在列於此。

表 3. 6 子資料集 2(Feature5,Feature7,Feature8)

	Feature5	Feature7	Feature8
A	4	4	10
B	4		2
C	4	4	2
D		5	1
E	4		2
F	5	5	2
G	6	5	

在本小節當中，主要是利用在第二章第二節當中，關於使用類神經網路來達到函式模擬的部分加以說明。目的是希望將所缺失的欄位值用填補之。下面將步驟方法以圖 3.4 說明之。

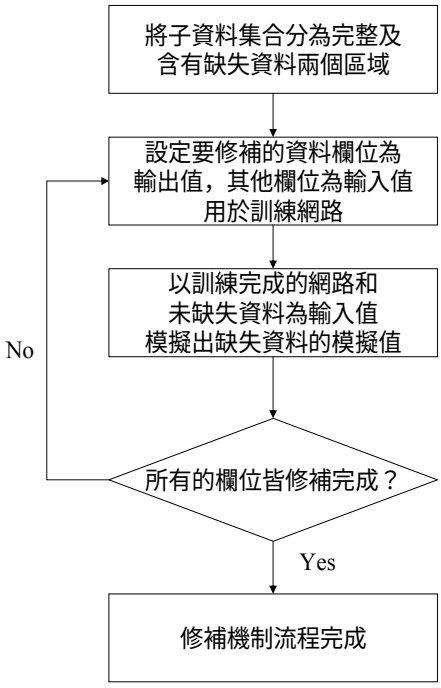


圖 3. 4 資料修補機制流程圖

以表 3.6 的例子來實際執行說明。

1. 根據 3.3.2 或是 3.3.3 的方法可以列出欲做出函式模擬的各特徵值項目子集合(Feature5、Feature7、Feature8)。
2. 將完整資料的部分集中，整理表 3.6 成為表 3.7 的模式。
3. 再將 Feature5 當作是類神經網路的輸出值，其餘特徵值項目（Feature7、Feature8）是做為類神經網路的輸入值，以初始的類神經網路加以訓練。
4. 再以訓練完成的類神經網路，以項目（Feature7、Feature8）為輸入值，使用類神經網路模擬的功能，將表 3.4 當中缺失的資料模擬出。
5. 將完整資料的部分集中，將表 3.6 整理為表 3.8 的模式。
6. 將 Feature7 當作是類神經網路的輸出值，其餘項目（Feature5、Feature8）是做為類神經網路的輸入值，以初始的類神經網路加以訓練。
7. 以訓練完成的類神經網路，以項目（Feature5、Feature8）為輸入值，使用類神經網路模擬的功能，將表 3.5 當中缺失的資料模擬出。
8. 將完整資料的部分集中，將表 3.6 整理為表 3.9 的模式。
9. 將 Feature8 當作是類神經網路的輸出值，其餘項目（Feature5、Feature7）是做為類神經網路的輸入值，以初始的類神經網路加以訓練。
10. 以訓練完成的類神經網路，以項目（Feature5、Feature7）為輸入值，使用類神經網路模擬的功能，將表 3.9 當中缺失的資料模擬出。
11. 資料修補結束。

表 3. 7 修補子集 1

	Feature5	Feature7	Feature8
A	4	4	10
C	4	4	2
F	5	5	2
D		5	1

表 3. 8 修補子集 2

	Feature5	Feature7	Feature8
A	4	4	10
C	4	4	2
F	5	5	2
B	4		2
E	4		2

表 3. 9 修補子集 3

	Feature5	Feature7	Feature8
A	4	4	10
C	4	4	2
F	5	5	2
G	6	5	

3.3.5 資料合併

經過 3.3.4 資料修補的過程之後，便根據缺失資料所修補的新值填入原始缺失資料子集當中。而形成表 3.10 的模式。

表 3. 10 缺失欄位修補結果

	Feature5	Feature7	Feature8
A	4	4	10
B	4	3.9999	2
C	4	4	2
D	4.8311	5	1
E	4	3.9999	2
F	5	5	2
G	6	5	2.1252

註：灰色部分表示以修補完成的特徵值

在根據由於 3.3.2 或是 3.3.3 所分解出來的子集回填入主要資料集，產生出表 3.11。

表 3. 11 表 3.1 之修補結果

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8
A	2	9	5	4.8514*	4	5	4	10
B	2		5	5	4		3.9999*	2
C	3	9	5	4.8645*	4		4	2
D	2	9	5	4	4.8311*		5	1
E	2		5	5	4		3.9999*	2
F	2	9	5.0014*	4	5		5	2
G	2	9	5	4	6		5	2.1252*

註：星號部分表示以修補完成的特徵值，灰色部分表示該特徵值為資料缺失的特徵值。

3.3.6 資料分類 – 以支援向量機器作為分類工具

本研究使用 LIBSVM 作為分類工具。下面將對 LIBSVM 的主程式部分及本研究所使用到的資料做個說明。

(1) 資料型態

LIBSVM 的檔案格式必須為如下

[label] [Index1]:[value1] [Index2]:[value2] ...

[label] [Index1]:[value1] [Index2]:[value2] ...

格式說明：

label 或稱之 class，分類的種類。

Index 為有順序的索引，通常是連續的整數。

value 就是用來訓練的資料內容值。

舉例來說，1 1:0 2:3 4:3。這表示為分類為 1，第一個特徵值值為 0，第二個特徵值值為 0，第三個特徵值值為 0。

(2) 主程式說明

svmtrain

訓練(train)過程當中會接受特定格式的輸入， 產生一個 "Model" 檔。這個 model 可以想像成 SVM 的內部資料，因為預測(predict)要 model 才能 predict，不能直接讀取原始資料。假定 train 本身是很耗時的動作，而訓練好可以以某種形式存起內部資料，那之後要預測時直接把那些內部資料讀取進來較為節省時間。

svmpredict

依照已經訓練完成的 model， 再加上給定的輸入 (新值)， 輸出預測(predict)新值所對應的類別 (class)。

(3)本研究使用 LIBSVM 說明

透過自行撰寫的轉換程式將 Matlab 中已經過修補的資料表轉換成 LIBSVM 能使用的格式。

將以修補後的資料分為兩群分別為訓練資料以及測試資料。將資料表以隨機方式選取三分之二資料量為訓練資料集，訓練類神經分類器；剩餘的三分之一資料量做為測試資料集，以評估整體架構的可應用性。

第四章 實例驗證

本章當中，將以三個完整的完整資料集來進行本研究方法的試驗步驟，用來測試本研究方法之合理性。最後新光醫院正子中心資料集(有著缺失資料之醫療資訊資料)來檢驗本方法應用在醫療資訊的是否適用。

4.1 IRIS Plants 資料集

IRIS 資料集最早由 R.A. Fisher 在 "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936)，一文當中提出來使用。希望藉由著收集且觀察花瓣及花萼本身的觀測值，辨識出為何種 IRIS 植物。下表 4. 1 為 IRIS Plants Database 的基本項目。

表 4. 1　IRIS Plants Database 基本資料表

資料集名稱	IRIS Plants Database
資料筆數	在本資料集當中有 3 類共 150 筆資料。
缺失情形	無缺失資料

4.1.1 資料特徵值說明

下表 4. 2 為 IRIS Plants Database 之特徵值項目說明。

表 4. 2 IRIS Plants Database 特徵值列表

順序	內容說明
1	sepal length
2	sepal width
3	petal length
4	petal width
5	該筆資料類別

4.1.2 資料前置分析及處理

原始資料檔案如下所示（節錄）：

IRIS.data

5.1,3.5,1.4,0.2,IRIS-setosa

4.9,3.0,1.4,0.2,IRIS-setosa

…

確定共有三類資料(IRIS Setosa、IRIS Versicolour、IRIS Virginica)。將該筆資料的類別轉換成數值型態，分別為 1,2,3 來作為表示。並對於各個特徵項目作初步之統計分析。下表 4. 3 為此資料庫的統計資料。

表 4. 3 IRIS 統計資料表

特徵項目	最小值	最大值	平均值	標準差
Sepal_length	4.3	7.9	5.84	0.83
Sepal_width	2.0	4.4	3.05	0.43
Petal_length	1.0	6.9	3.76	1.76
Petal_width	0.1	2.5	1.20	0.76

4.1.3 試驗說明

本研究實驗當中，將有二個不同的實驗因子來進行實驗說明。

第一個實驗因子為缺失情形

將以缺失率為 0.1、0.5、0.9 來說明資料缺失比例對於修補機制的影響。

第二個實驗因子為類神經網路架構

將以不同之網路架構為來說明類神經網路架構對於修補機制的影響。對於修補機制的影響。

試驗步驟說明：

第一步驟 - 實驗前設定：

根據本研究第三章的研究方法。首先要選定特徵項目之間的關連性。由計算交互資訊程式計算出的交互資料表，見表 4. 4。

表 4. 4 IRIS Plants Database 交互資料表

	1	2	3	4	5
1	-	0.41242	0.15311	0.85828	0.88319
2	0.41242	-	0.16179	0.59619	0.498
3	0.15311	0.16179	-	0.23383	0.2796
4	0.85828	0.59619	0.23383	-	0.88259
5	0.88319	0.498	0.2796	0.88259	-

以及根據 MI 門檻值為 0.1 所選定的特徵項目關聯表(表 4. 5)。

所選定出來的相關特徵的關係分別為：

1. Sepal_length_{ Sepal_width , Petal_length , Petal_width }
2. Sepal_width_{ Sepal_length , Petal_length , Petal_width }
3. Petal_length _{ Sepal_length , Sepal_width , Petal_width }
4. Petal_width_{ Sepal_length , Sepal_width , Petal_length }

表 4. 5 特徵項目關聯表

欄位代碼	相關欄位代碼			
2	3	4	5	
3	2	4	5	
4	2	3	5	
5	2	3	4	

根據 3.3.4 的研究方法，各產生屬於該欄位的訓練資料集以及模擬資料集兩部分。

第二步驟：

設定下列三種狀態，試著想表達資料缺失的情境，共分為低度缺失、中度缺失、高度缺失的現象。

設定缺失率為 0.1，即為原始資料發生缺失情形的機會為 10%。

設定缺失率為 0.5，即為原始資料發生缺失情形的機會為 50%。

設定缺失率為 0.9，即為原始資料發生缺失情形的機會為 90%。

將原始完整的資料根據上述所設定的機率產生缺失現象。

第三步驟 - 設定類神經網路架構：

根據上步驟所選取出來的特徵項目關連當作類神經的輸入及輸出變數。建構出四個類神經網路來分別來處理函式模擬。下面流程為以缺失比率為 0.9，且類神經網路架構為單階隱藏層模型架構的一個範例流程說明。

(1) IRIS-類神經網路-1

根據 MI 表所選取出來的 Sepal_length_{ Sepal_width , Petal_length , Petal_width }，表示 Sepal_length 該欄位為目標輸出變數，Sepal_width , Petal_length , Petal_width 為輸入變數，用以訓練類神經分類器。

本研究應用 MATLAB 6.5 版建立與訓練單階隱藏層模型 (Single Hidden Layer Model)。進行網路架構之訓練與模擬，隱藏元之個數設為 10 個，類神經架構見下圖 4. 1。

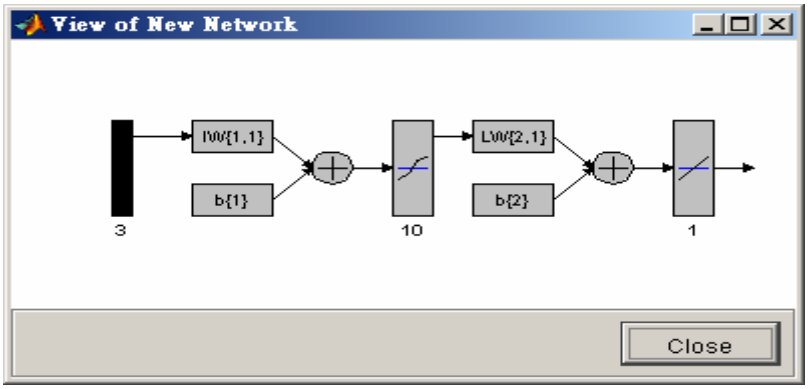


圖 4. 1 IRIS-1 類神經架構圖

以 3 個輸入變數、10 個隱藏元及 1 個輸出變數。建立網路模型，選定隱藏元之內部之活動函數為 Hyperbolic tangle function，而輸出元為 pure linear function。初始化網路權數值 (weights)之後，將完整資料集輸入訓練網路。網路穩定之後，將缺失該欄位(Sepal_length)模擬資料集開始模擬，取得所模擬之值補上缺失欄位。下圖表示網路已達到所設定之 Performance。

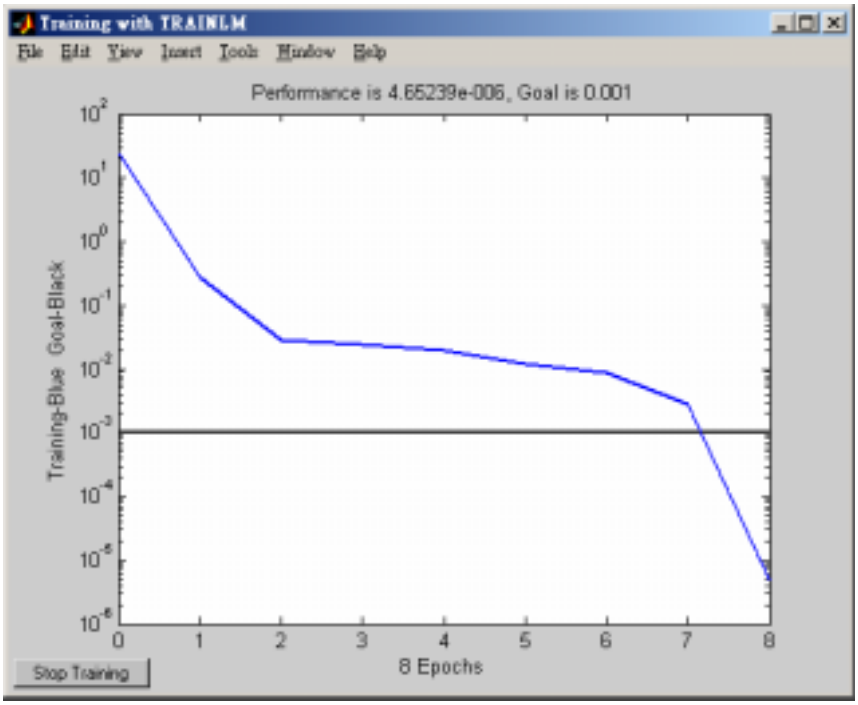


圖 4. 2 IRIS-類神經網路-1

(2) IRIS-類神經網路-2

根據 MI 表所選取出來的 Sepal_width_{ Sepal_length , Petal_length , Petal_width }，表示 Sepal_width 該欄位為目標輸出變數，Sepal_length , Petal_length , Petal_width 為輸入變數，用以訓練類神經分類器。

同樣的建立與訓練單階隱藏層模型 (Single Hidden Layer model)。進行網路架構之訓練與模擬，隱藏元之個數設為 10 個，類神經架構與 IRIS-類神經網路-1 相同。

如同(1)的網路架構一樣以 3 個輸入變數 10 個隱藏元及 1 個輸出變數。建立網路模型，選定隱藏元之內部之活動函數為 Hyperbolic

tangle function，而輸出元為 pure linear function。初始化網路權數值(weights)之後，將完整資料集輸入訓練網路。網路穩定之後，將缺失該欄位(Sepal_width)模擬資料集開始模擬，取得所模擬之值補上缺失欄位。圖 4. 3 表示經過了 190epoch，網路已達到所設定之 Performance。

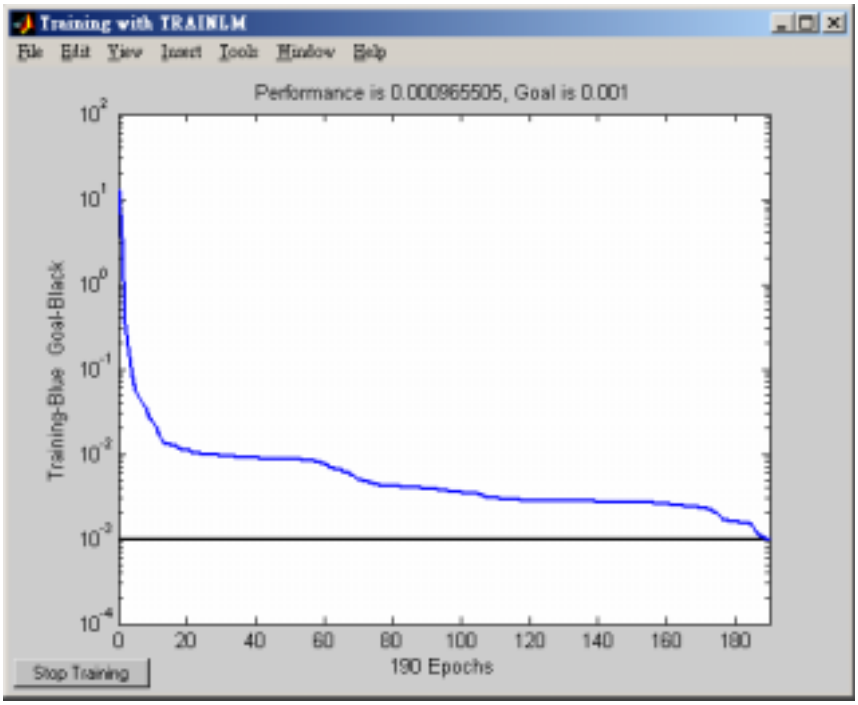


圖 4. 3 IRIS-類神經網路-2

(3) IRIS-類神經網路-3

根據 MI 表所選取出來的 Petal_length_{ Sepal_length , Sepal_width , Petal_width }，表示 Petal_length 該欄位為目標輸出變數，Sepal_length , Sepal_width , Petal_width 為輸入變數，用以訓練類神經分類器。

同樣的建立與訓練單階隱藏層模型 (Single Hidden Layer model)。進行網路架構之訓練與模擬，隱藏元之個數設為 10 個，類神經架構與 IRIS-類神經網路-1 相同。

如同(1)的網路架構一樣以 3 個輸入變數、10 個隱藏元及 1 個輸出變數。建立網路模型，選定隱藏元之內部之活動函數為 Hyperbolic tangle function，而輸出元為 pure linear function。初始化網路權數值(weights)之後，將完整資料集輸入訓練網路。網路穩定之後，將缺失

該欄位(Petal_length)模擬資料集開始模擬，取得所模擬之值補上缺失欄位。圖 4. 4 表示經過了 1000epoch，網路已達到所設定之 epoch 上限。

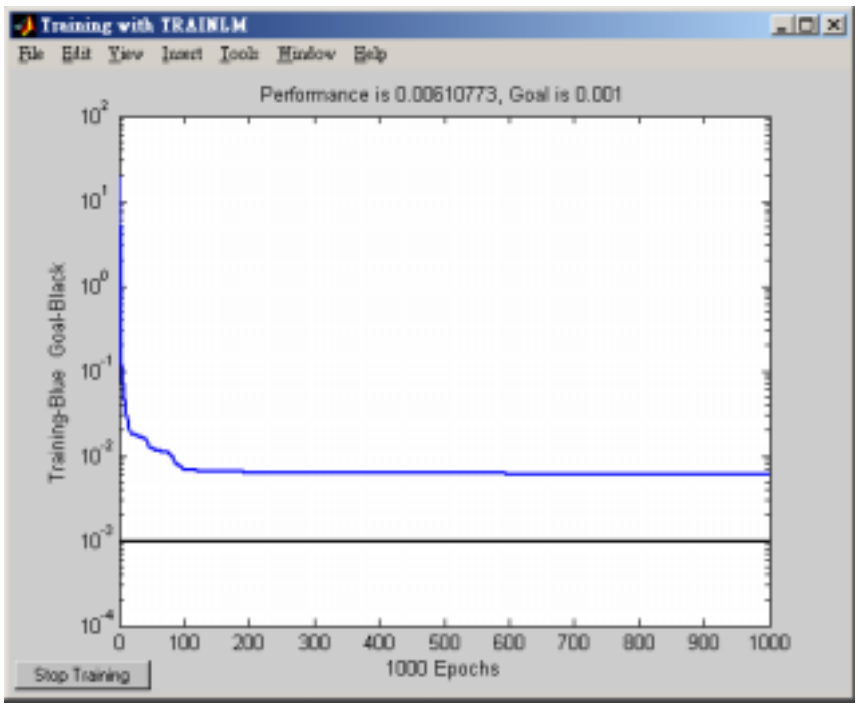


圖 4. 4 IRIS-類神經網路-3

(4) IRIS-類神經網路-4

根據 MI 表所選取出來的 Petal_width_{ Sepal_length , Sepal_width , Petal_length }，表示 Petal_width 該欄位為目標輸出變數，Sepal_length , Sepal_width , Petal_length 為輸入變數，用以訓練類神經分類器。

同樣的建立與訓練單階隱藏層模型 (Single Hidden Layer model)。進行網路架構之訓練與模擬，隱藏元之個數設為 10 個，類神經架構與 IRIS-類神經網路-1 相同。

如同(1)的網路架構一樣以 3 個輸入變數 10 個隱藏元及 1 個輸出變數。建立網路模型，選定隱藏元之內部之活動函數為 Hyperbolic tangle function，而輸出元為 pure linear function。初始化網路權數值(weights)之後，將完整資料集輸入訓練網路。網路穩定之後，將缺失該欄位(Petal_length)模擬資料集開始模擬，取得所模擬之值補上缺失

欄位。圖 4. 5 表示經過了 1000epoch，網路已達到所設定之 epoch 上限。

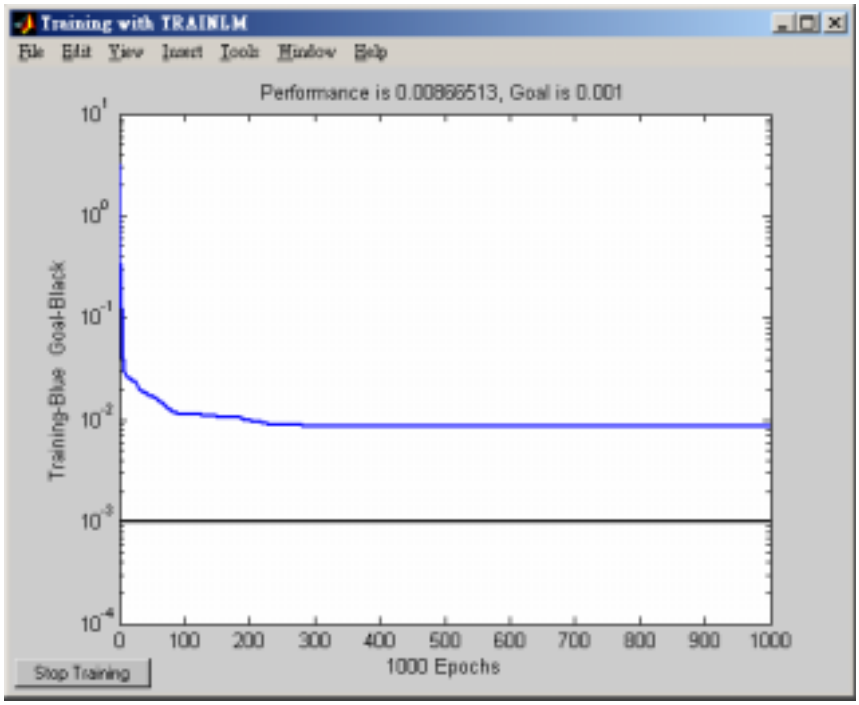


圖 4. 5 IRIS-類神經網路-4

4.1.4 試驗結果紀錄

根據以上的設定，試驗了在不同的缺失情形及類神經網路架構下，對於修補過後的資料來做分類的試驗。結果如下表，原始資料正確性為使用完整資料做分類的結果，缺失資料正確率為使用缺失資料作分類之結果，修補過後正確率為使用修補機制過後的資料做分類之結果。

下面將以三種缺失值（0.1、0.5、0.9）以及三種類神經網路架構分別為單階隱藏層模型-10 神經元（圖 4.6）、單階隱藏層模型-50 神經元（圖 4.7）、雙階隱藏層模型-10 神經元、10 神經元（圖 4.8），來做資料修補機制的實驗控制變數。

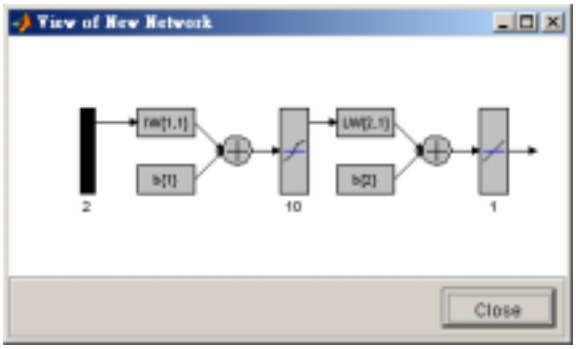


圖 4. 6 單階隱藏層模型-10 神經元

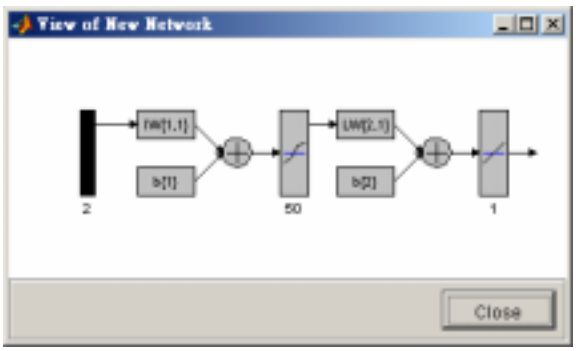


圖 4. 7 單階隱藏層模型-50 神經元

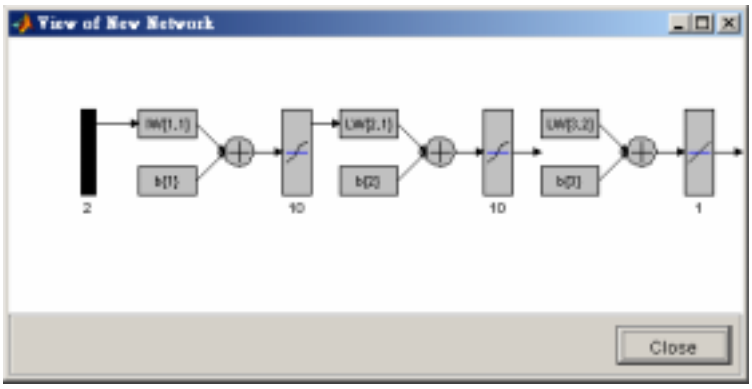


圖 4. 8 雙階隱藏層模型-10 神經元、10 神經元

試驗一(類神經網路架構為單階隱藏層模型-10 神經元)：

共 51 筆測試資料

缺失比率	90%	50%	10%
缺失資料表經 SVM 分類測試正確個數	42	46	48
	45	46	49
	43	43	47
	44	45	48
	41	42	48
	49	44	48
	46	43	47
	45	41	49
	40	43	46
	42	41	45
修補過後資料表經 SVM 分類測試正確個數	48	47	48
	48	48	50
	44	45	47
	47	49	50
	45	47	48
	50	46	48
	47	45	48
	46	46	49
	45	47	47
	43	43	49

試驗二(類神經網路架構為單階隱藏層模型-50 神經元)：

共 51 筆測試資料

缺失比率	90%	50%	10%
缺失資料表經 SVM 分類測試正確個數	44	46	47
	45	45	49
	44	42	49
	42	47	48
	40	45	46
	44	47	48
	47	48	46
	43	45	45
	41	48	47
	43	45	48
修補過後資料表經 SVM 分類測試正確個數	46	48	50
	47	48	49
	45	49	50
	47	49	47
	45	47	48
	46	50	48
	47	48	48
	43	48	46
	46	47	49
	43	47	49

試驗三(類神經網路架構為雙階隱藏層模型-10 神經元、10 神經元)：
共 51 筆測試資料

缺失比率	90%	50%	10%
缺失資料表經 SVM 分類測試正確個數	44	46	46
	45	44	47
	45	44	48
	47	43	47
	44	44	48
	44	47	47
	48	45	48
	42	43	47
	47	46	48
	44	46	47
修補過後資料表經 SVM 分類測試正確個數	47	49	45
	46	49	49
	46	49	49
	44	48	49
	44	47	49
	44	47	48
	48	49	46
	46	45	49
	49	48	50
	48	47	49

4.1.5 IRIS Plants 資料集- 試驗結論

本試驗結果將以重複量數統計法(三因子變異數分析)來做為驗證工具。以下是本研究統計分析的設定。

A 因子為分別缺失狀態時分類結果及經過修補機制後分類結果。

B 因子為缺失情形的設定，共有三個水準(90%,50%,10%的缺失情形)。

C 因子為類神經網路的設定，共有三個水準(單階隱藏層模型-10 神經元，單階隱藏層模型-50 神經元，雙階隱藏層模型-10 神經元、10 神經元)。

在本研究欲希望得知，經過修補機制過後，正確率是否有顯著的差異，並不去探討在不同缺失情形及不同類神經網路的情境下之間的差異，故在此僅說明 A 因子（經過修補機制）之結果。在這裡使用的統計工具為 SPSS11.0 for windows 版本用來計算其結果，表 4.6 為程式執行結果。

其虛無假設 H_0 為 $\mu_{修補前} = \mu_{修補後}$，其對立假設 H_1 為 $\mu_{修補前} \neq \mu_{修補後}$

以顯著水準 α 為 0.05 來檢定，計算出其 F 值為 $61.89138 > F_{(0.95;1,163)} = 3.86$

在此否定其虛無假設 H_0，即為兩組數據之間是有差異的。

A 因子為經過修補機制過後，有著顯著的差異。根據各平均值結果(修補前為 45.344，修補過後 47.256)亦可以說明修補過後的資料集經過分類結果正確個數高於未修補時的資料集。

表 4.6 修補前後之變異數分析表

Source	SS	df	MS	F	Sig.
A	164.356	1	164.356	61.89138	0.000
B	12.933	2	6.4665	2.435084	0.091
C	252.933	2	126.4665	47.62337	0.000
A*B	1.244	2	0.622	0.234226	0.791
A*C	22.711	2	11.3555	4.27613	0.015
B*C	47.933	4	11.98325	4.512521	0.002
A*B*C	5.489	4	1.37225	0.516747	0.724
Error	430.2	163	2.655556		
Total	937.8	180			

4.2 Glass Identification 資料集

此資料集為對進行現場的玻璃碎片進行成分分析。希望藉由玻璃裡面元素含量值來辨識為何種原始型態的物品。預期訓練分類機制能達到辨識該原始型態。下表 4. 7 Glass Identification 基本資料表為此資料集之基本資料。

表 4. 7 Glass Identification 基本資料表

資料集名稱	Glass Identification
資料筆數	在本資料集當中有 7 類共 214 筆資料。
缺失情形	無缺失資料

4.2.1 資料特徵值說明

下表 4. 8 為 Glass Identification 之特徵值項目說明。

表 4. 8 Glass Identification 特徵值列表

順序	內容說明
1	Id number
2	refractive index
3	Sodium
4	Magnesium
5	Aluminum
6	Silicon
7	Potassium
8	Calcium
9	Barium
10	Iron
11	Type of glass

4.2.2 資料前置分析及處理

原始資料檔案如下所示（節錄）：

1,1.52101,13.64,4.49,1.10,71.78,0.06,8.75,0.00,0.00,1

2,1.51761,13.89,3.60,1.36,72.73,0.48,7.83,0.00,0.00,1

3,1.51618,13.53,3.55,1.54,72.99,0.39,7.78,0.00,0.00,1

4,1.51766,13.21,3.69,1.29,72.61,0.57,8.22,0.00,0.00,1

…

確定共有 7 類資料(1,2,…,7)。由於資料本身以做轉換，故不用作轉換的動作。第一欄位為識別碼，對於分析並不具有任何意義，在此刪除之。

表 4. 9 Glass Identification 統計資料

特徵項目	最小值	最大值	平均值	標準差
refractive index	1.51115	1.53393	1.518353	0.003039
Sodium	10.73	17.38	13.4068	0.81837
Magnesium	0	4	2.676	1.4405
Aluminum	0.29	3.5	1.4465	0.49988
Silicon	69.81	75.41	72.655	0.77405
Potassium	0	6.21	0.4991	0.65304
Calcium	5.43	16.19	8.9579	1.42643
Barium	0	3	0.18	0.498
Iron	0	1	0.06	0.098

4.2.3 試驗說明

本研究實驗當中，將有二個不同的實驗因子來進行實驗說明。

第一個實驗因子為缺失情形

將以缺失率為 0.1、0.5、0.9 來說明資料缺失比例對於修補機制的影響。

第二個實驗因子為類神經網路架構

將以不同之網路架構為來說明類神經網路架構對於修補機制的影響。對於修補機制的影響。

試驗步驟說明：

第一步驟 - 實驗前設定：

根據本研究第三章的研究方法。首先要選定特徵項目之間的關連性。由計算交互資訊程式計算出的交互資料表(表 4.8)。以及根據 MI 門檻值為 0.1 所選定的特徵項目關聯表(表 4.12)，所選定出來的相關特徵的關係分別如表 4.10。

根據 3.3.4 的研究方法，各產生屬於該欄位的訓練資料集以及模擬資料集兩部分以供下步驟使用。

第二步驟：

設定下列三種狀態，試著想表達資料缺失的情境，共分為低度缺失、中度缺失、高度缺失的現象。

設定缺失率為 0.1，即為原始資料發生缺失情形的機會為 10%。

設定缺失率為 0.5，即為原始資料發生缺失情形的機會為 50%。

設定缺失率為 0.9，即為原始資料發生缺失情形的機會為 90%。

將原始完整的資料根據上述所設定的機率產生缺失現象。

表 4. 10 Glass Identification 交互資料表-1

	1	2	3	4	5
1	-	0.090571	0.19352	0.3892	0.26002
2	0.090571	-	0.11411	0.07507	0.11532
3	0.19352	0.11411	-	0.20818	0.11715
4	0.3892	0.07507	0.20818	-	0.17671
5	0.26002	0.11532	0.11715	0.17671	-
6	0.04593	0.14462	0.089513	0.15844	0.08024
7	-0.00347	-0.03271	-0.05016	-0.05658	-0.01476
8	0.18682	0.31466	0.085697	0.20166	0.13694
9	0.24824	-0.002	0.094378	0.094188	0.1948
10	0.041206	-0.00402	0.051572	0.033388	-0.03448

表 4. 11 Glass Identification 交互資料表-2

	6	7	8	9	10
1	0.04593	-0.00347	0.18682	0.24824	0.041206
2	0.14462	-0.03271	0.31466	-0.002	-0.00402
3	0.089513	-0.05016	0.085697	0.094378	0.051572
4	0.15844	-0.05658	0.20166	0.094188	0.033388
5	0.08024	-0.01476	0.13694	0.1948	-0.03448
6	-	0.002319	0.14669	0.013888	-0.00064
7	0.002319	-	0.062372	0.001488	-0.06017
8	0.14669	0.062372	-	0.012789	0.03436
9	0.013888	0.001488	0.012789	-	-0.01879
10	-0.00064	-0.06017	0.03436	-0.01879	-

表 4. 12 Glass Identification 特徵項目關聯表(僅代碼)

欄位代碼	相關欄位代碼							
2	3	4	5	6	8			
3	2	4	5	6	8	9	10	
4	2	3	5	6	8	9		
5	2	3	4	6	8	9		
6	2	3	4	5	8			
7	8							
8	2	3	4	5	6	7		
9	3	4	5					
10	3							

第三步驟 - 設定類神經網路架構：

根據上步驟所選取出來的特徵項目關連當作類神經的輸入及輸出變數。建構共 9 個類神經網路來分別來處理函式模擬。下面流程為以缺失比率為 0.9，且類神經網路架構為單階隱藏層模型架構的一個範例流程說明。

(1) Glass Identification-類神經網路-1 (舉例說明)

根據 MI 表所選取出來的 Refractive_index_{ Sodium, Magnesium, Aluminum, Silicon, Calcium }，表示 Refractive_index 該欄位為目標輸出變數，odium, Magnesium, Aluminum, Silicon, Calcium 為輸入變數，用以訓練類神經分類器。

本研究應用 MATLAB 6.5 版建立與訓練單階隱藏層模型 (Single Hidden Layer model)。進行網路架構之訓練與模擬，隱藏元之個數設為 10 個，類神經架構見圖 4.9。

以 5 個輸入變數、10 個隱藏元及 1 個輸出變數。建立網路模型，選定隱藏元之內部之活動函數為 Hyperbolic tangle function，而輸出元為 pure linear function。初始化網路權數值 (weights)之後，將完整資

料集輸入訓練網路。網路穩定之後，將缺失該欄位(Refractive_index)模擬資料集開始模擬，取得所模擬之值補上缺失欄位。

(2)-(10) Glass Identification-類神經網路-2-9

由於建構類神經網路的過程大致相同，因此 Glass Identification-類神經網路-2 至 Glass Identification-類神經網路-9 的建構部分將僅敘述。

根據 MI 表所選取出來的預修補的欄位為目標輸出變數，相關欄位（個數不一定，由計算出來的結果來取代）為輸入變數，用以訓練類神經分類器。

本研究應用 MATLAB 6.5 版建立與訓練單階隱藏層模型 (Single Hidden Layer model)。進行網路架構之訓練與模擬，隱藏元之個數設為 10 個，類神經架構見圖 4.9。

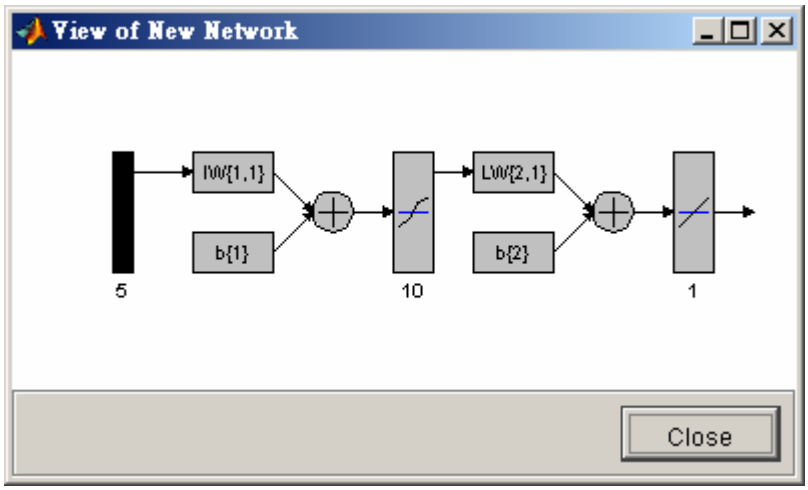


圖 4. 9 Glass-類神經架構圖-1

表 4. 13 Glass Identification 特徵項目關聯表(欄位名稱)

欄位代碼	預修補的欄位	相關欄位
2	Refractive index	Sodium, Magnesium, Aluminum, Silicon, Calcium
3	Sodium	Refractive index, Magnesium, Aluminum, Silicon, Calcium, Calcium, Barium, Iron
4	Magnesium	Refractive index, Sodium, Aluminum, Silicon, Calcium, Calcium, Barium
5	Aluminum	Refractive index, Sodium, Magnesium, Silicon, Calcium, Calcium, Barium
6	Silicon	Refractive index, Sodium, Magnesium, Aluminum, Calcium, Calcium, Barium
7	Potassium	Calcium
8	Calcium	Refractive index, Sodium, Magnesium, Aluminum, Silicon ,Potassium,
9	Barium	Sodium, Magnesium, Aluminum
10	Iron	Sodium

以相關欄位個數為輸入變數、10 個隱藏元及 1 個輸出變數。建立網路模型，選定隱藏元之內部之活動函數為 Hyperbolic tangle function，而輸出元為 pure linear function。初始化網路權數值 (weights) 之後，將完整資料集輸入訓練網路。

網路穩定之後，將缺失該欄位模擬資料集開始模擬，取得所模擬之值補上缺失欄位。

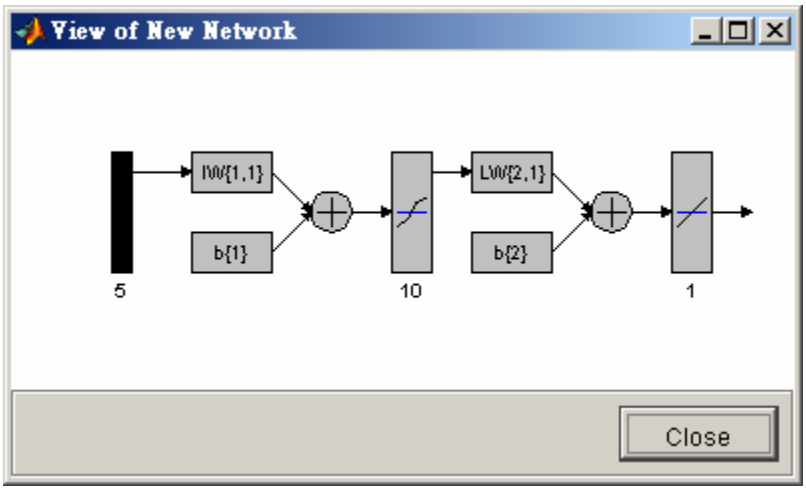


圖 4. 10 Glass-類神經架構圖-2

將個別類神經網路架構參數表列如下：

網路代碼	A(輸入變數個數)
2	7
3	6
4	6
5	5
6	1
7	6
8	3
9	1

4.2.4 試驗結果紀錄

根據以上的設定，試驗了在不同的缺失情形及類神經網路架構下，對於修補過後的資料來做分類的試驗。結果如下表，原始資料正確性為使用完整資料做分類的結果，缺失資料正確率為使用缺失資料作分類之結果，修補過後正確率為使用修補機制過後的資料做分類之結果。

下面將以三種缺失值（0.1、0.5、0.9）以及三種類神經網路架構分別為單階隱藏層模型-10 神經元（圖 4.6）、單階隱藏層模型-50 神經元（圖 4.7）、雙階隱藏層模型-10 神經元、10 神經元（圖 4.8），來做資料修補機制的實驗控制變數。

試驗一(類神經網路架構為單階隱藏層模型-10 神經元)：

共 75 筆測試資料

缺失比率	90%		50%		10%	
缺失資料表經 SVM 分類測試正確個數	34	48	36	47	47	49
	37	46	40	46	48	49
	41	42	39	50	47	48
	39	51	39	49	43	47
	38	47	42	47	48	50
修補過後資料表經 SVM 分類測試正確個數	34	44	40	50	47	48
	35	46	38	48	47	49
	38	42	37	46	46	49
	33	49	39	47	48	49
	36	45	38	39	47	48

試驗二(類神經網路架構為單階隱藏層模型-50 神經元)：

共 75 筆測試資料

缺失比率	90%		50%		10%	
缺失資料表經 SVM 分類測試正確個數	39	48	40	48	44	48
	32	46	45	46	49	48
	39	41	45	46	45	49
	36	50	40	43	48	50
	36	43	42	46	47	49
修補過後資料表經 SVM 分類測試正確個數	38	46	43	47	48	47
	35	46	45	47	46	50
	32	40	43	46	45	49
	37	41	42	46	45	48
	36	44	43	46	46	47

試驗三(類神經網路架構為雙階隱藏層模型-10 神經元、10 神經元)：

共 75 筆測試資料

缺失比率	90%		50%		10%	
缺失資料表經 SVM 分類測試正確個數	32	44	37	48	42	48
	36	47	36	42	46	43
	35	45	39	46	48	44
	32	50	40	42	43	49
	34	46	42	46	47	50
修補過後資料表經 SVM 分類測試正確個數	38	48	38	42	46	50
	38	48	42	43	42	49
	33	45	43	45	39	48
	37	48	47	46	43	46
	35	49	39	41	44	47

4.2.5 Glass Identification Database 資料集- 試驗結論

本試驗結果將以重複量數統計法(三因子變異數分析)來做為驗證工具。以下是本研究統計分析的設定。

A 因子為分別缺失狀態時分類結果及經過修補機制後分類結果。

B 因子為缺失情形的設定，共有三個水準(90%,50%,10%的缺失情形)。

C 因子為類神經網路的設定，共有三個水準(單階隱藏層模型-10 神經元，單階隱藏層模型-50 神經元，雙階隱藏層模型-10 神經元、10 神經元)。

在本欲希望得知，經過修補機制過後，正確率是否有顯著的差異，並不去探討在不同缺失情形及不同類神經網路的情境下之間的差異，故在此僅說明 A 因子（經過修補機制）之結果。在這裡使用的統計工具為 SPSS11.0 for windows 版本用來計算其結果，表 4. 14 為程式執行結果。

其虛無假設 H_0 為 $\mu_{修補前} = \mu_{修補後}$，其對立假設 H_1 為 $\mu_{修補前} \neq \mu_{修補後}$ 以顯著水準 α 為 0.05 來檢定，計算出其 F 值為 292.6933 > $F_{(0.95;1,163)} = 3.86$

在此否定其虛無假設 H_0，即為兩組數據之間是有差異的。

A 因子為經過修補機制過後，有著顯著的差異。根據各平均值結果(修補前為 40.722，修補過後 46.576)亦可以說明修補過後的資料集經過分類結果正確個數高於未修補時的資料集。

表 4. 14 修補前後之變異數分析表(Glass identification 試驗)

Source	SS	df	MS	F	Sig.
A	1537.089	1	1537.089	292.6933	0.000
B	41.64444	2	20.82222	3.964979	0.021
C	1136.844	2	568.4222	108.2393	0.000
A*B	31.24444	2	15.62222	2.974792	0.054
A*C	439.2444	2	219.6222	41.82059	0.000
B*C	69.35556	4	17.33889	3.301681	0.013
A*B*C	77.82222	4	19.45556	3.704738	0.006
Error	856	163	5.251534		
Total	4189.244	180			

4.3 Letter Image Recognition 資料集

此資料集收集目標為在黑白的矩形像素圖片上之 26 個大寫字母。所產生的特性圖像基於 20 個不同字型，並且由這 20 個字型之內的字母隨機產生了 20000 筆資料並記錄。預期訓練分類機制能達到辨識該字母。下表為此資料集之基本資料。由於本研究並不用到所有的資料，僅取十分之一的資料作為本次試驗的原始資料。

表 4. 15 Letter Image Recognition 基本資料表

資料集名稱	Letter Image Recognition
資料筆數	在本資料集當中有 26 類共 1987 筆資料。
缺失情形	無缺失資料

4.3.1 資料特徵值說明

下表 4. 16 為 Letter Image Recognition 之特徵值項目說明。

表 4. 16 Letter Image Recognition 特徵值項目

順序	欄位名稱	順序	欄位名稱
1.	letter	10.	y2bar
2.	x-box	11.	xybar
3.	y-box	12.	x2ybr
4.	width	13.	xy2br
5.	high	14.	x-ege
6.	onpix	15	xegvy
7.	x-bar	16	y-ege
8.	y-bar	17	yegvx
9.	x2bar		

4.3.2 資料前置分析及處理

原始資料檔案如下所示（節錄）：

LETTER-RECOGNITION.DATA

T,2,8,3,5,1,8,13,0,6,6,10,8,0,8,0,8

I,5,12,3,7,2,10,5,5,4,13,3,9,2,8,4,10

D,4,11,6,8,6,10,6,2,6,10,3,7,3,7,3,9

N,7,11,6,6,3,5,9,4,6,4,4,10,6,10,2,8

…

確定共有 26 類資料(A,B,C, ... ,Z)。將該筆資料的類別轉換成數值型態，分別為 1,2,3,…,26 來作為表示。

表 4. 17 Letter Image Recognition 統計資料表

特徵項目	最小值	最大值	平均值	標準差
x-box	0	15	7.0355	3.30456
y-box	0	15	5.1218	2.01457
width	0	15	5.3725	2.26139
high	0	15	3.5058	2.19046
onpix	0	15	6.8976	2.02604
x-bar	0	15	7.5005	2.32535
y-bar	0	15	4.6286	2.69997
x2bar	0	15	5.1786	2.38082
y2bar	0	15	8.282	2.48847
xybar	0	15	6.454	2.63107
x2ybr	0	15	7.929	2.08062
xy2br	0	15	3.0461	2.33254
x-ege	0	15	8.3388	1.54672
xegvy	0	15	3.6917	2.56707
y-ege	0	15	7.8012	1.61747
yegvx	0	15	7.0355	3.30456

4.3.3 試驗說明

本研究實驗當中，將有二個不同的實驗因子來進行實驗說明。

第一個實驗因子為缺失情形

將以缺失率為 0.1、0.5、0.9 來說明資料缺失比例對於修補機制的影響。

第二個實驗因子為類神經網路架構

將以不同之網路架構為來說明類神經網路架構對於修補機制的影響。對於修補機制的影響。

試驗步驟說明：

第一步驟 - 實驗前設定：

根據本研究第三章的研究方法。首先要選定特徵項目之間的關連性。由計算交互資訊程式得出的交互資料表，見表 4. 18。以及根據交互資訊門檻值為 0.1 所選定的特徵項目關聯表。

所選定出來的相關特徵的關係分別為：

表 4. 18 Letter Image Recognition 特徵項目關聯表

欄位代碼	預修補的欄位	相關欄位
2	x-box	y-box, width, high, onpix, x-ege
3	y-box	x-box, width, high, onpix
4	width	x-box, y-box, high, onpix, x-ege
5	high	x-box ,y-box, width , onpix
6	onpix	x-box ,y-box, width , high, x-ege , y-ege
7	x-bar	y-bar, xybar, x2ybr
8	y-bar	x-bar, xybar, x2ybr, xy2br, x-ege, xegvy, y-ege
9	x2bar	y2bar, xybar, xybar, x2ybr, x-ege, y-ege,
10	y2bar	x2bar, xybar, xybar, x-ege, y-ege
11	Xybar	x-bar, y-bar, x2bar, y2bar, x2ybr, x-ege, y-ege,

表 4. 19 Letter Image Recognition 特徵項目關聯表(續上頁)

12	x2ybr	x-bar, y-bar, x2bar, y2bar, xybar, x-ege, y-ege, yegvx
13	xy2br	y-bar
14	x-ege	x-box, width, onpix, y-bar, x2bar, y2bar, Xybar, x2ybr, xegvy, y-ege
15	xegvy	x2bar, x2ybr, x-ege
16	y-ege	onpix, y-bar, x2bar, y2bar, Xybar, x2ybr, xegvy, yegvx
17	yegvx	x2ybr, y-ege

根據 3.3.4 的研究方法，各產生屬於該欄位的訓練資料集以及模擬資料集兩部分以供下步驟使用。

表 4. 20　Letter Image Recognition 特徵項目關聯表(僅代碼)

修補代碼	相關欄位代碼										
2	3	4	5	6	14						
3	2	4	5	6							
4	2	3	5	6	14						
5	2	3	4	6							
6	2	3	4	5	14	16					
7	8	11	12								
8	7	11	12	13	14	15	16				
9	10	11	12	14	16						
10	9	11	14	16							
11	7	8	9	10	12	14	16				
12	7	8	9	11	14	15	16	17			

表 4. 21　Letter Image Recognition 特徵項目關聯表(續上頁)

13	8										
14	2	4	6	8	9	10	11	12	15	16	
15	8	12	14								
16	6	8	9	10	11	12	14	17			
17	12	16									

第二步驟：

設定下列三種狀態，試著想表達資料缺失的情境，共分為低度缺失、中度缺失、高度缺失的現象。

設定缺失率為 0.1，即為原始資料發生缺失情形的機會為 10%。

設定缺失率為 0.5，即為原始資料發生缺失情形的機會為 50%。

設定缺失率為 0.9，即為原始資料發生缺失情形的機會為 90%。

將原始完整的資料根據上述所設定的機率產生缺失現象。

第三步驟 - 設定類神經網路架構：

根據上步驟所選取出來的特徵項目關連當作類神經的輸入及輸出變數。建構共 16 個類神經網路來分別來處理函式模擬。下面流程為以缺失比率為 0.9，且類神經網路架構為單階隱藏層模型架構的一個範例流程說明。

(1) Letter Image Recognition-類神經網路-1 (舉例說明)

根據 MI 表所選取出來的 x-box _{ y-box, width, high, onpix, x-ege }，表示 x-box 該欄位為目標輸出變數，y-box, width, high, onpix, x-ege 為輸入變數，用以訓練類神經分類器。

本研究應用 MATLAB 6.5 版建立與訓練單階隱藏層模型 (Single Hidden Layer model)。進行網路架構之訓練與模擬，隱藏元之個數設為 10 類神經架構見下圖 4. 11。

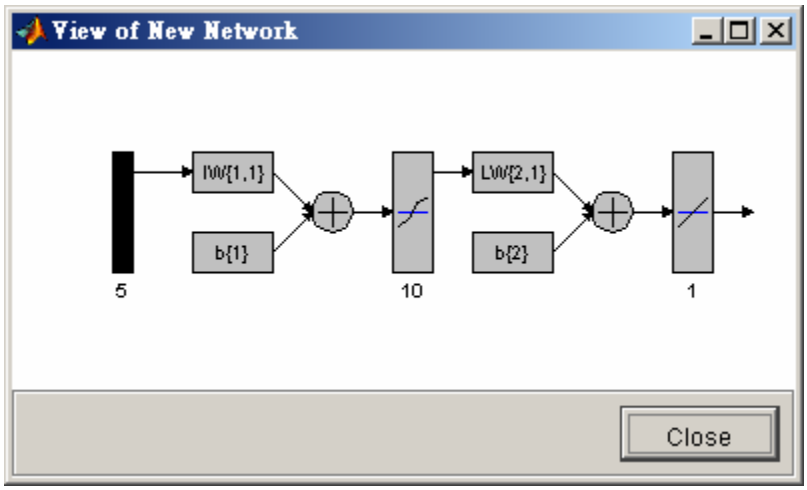


圖 4. 11 LETTER-類神經架構圖-1

以 5 輸入變數、10 個隱藏元及 1 個輸出變數。建立網路模型，選定隱藏元之內部之活動函數為 Hyperbolic tangle function，而輸出元為 pure linear function。初始化網路權數值 (weights)之後，將完整資料集輸入訓練網路。網路穩定之後，將缺失該欄位(x-box)模擬資料集開始模擬，取得所模擬之值補上缺失欄位。

(2)-(17) Letter Image Recognition-類神經網路-2-16

由於建構類神經的過程大致相同，所以根據 Letter Image Recognition-類神經網路-2 至 Letter Image Recognition-類神經網路-16 的建構部分將僅擇一敘述，表來表示其收斂情形。

根據 MI 表所選取出來的預修補的欄位為目標輸出變數，相關欄位（個數不一定，由計算出來的結果來設定，見表 4. 18）為輸入變數，用以訓練類神經分類器。

本研究應用 MATLAB 6.5 版建立與訓練單階隱藏層模型 (Single Hidden Layer model)。進行網路架構之訓練與模擬，隱藏元之個數設為 10 個，類神經架構見下圖 4. 12。

以相關欄位個數為輸入變數、10 個隱藏元及 1 個輸出變數。建立網路模型，選定隱藏元之內部之活動函數為 Hyperbolic tangle

function，而輸出元為 pure linear function。初始化網路權數值 (weights) 之後，將完整資料集輸入訓練網路。網路穩定之後，將缺失該欄位模擬資料集開始模擬，取得所模擬之值補上缺失欄位。

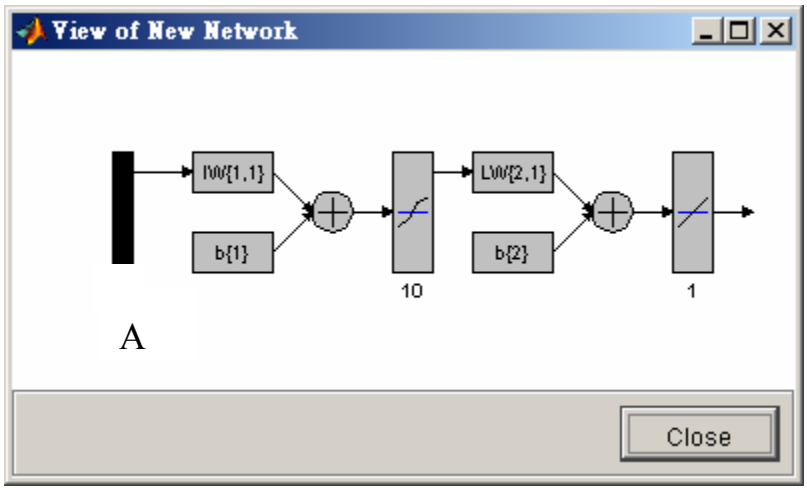


圖 4. 12 LETTER-類神經架構圖-2

將個別類神經網路架構參數表列如下：

網路代碼	A(輸入變數個數)	網路代碼	A(輸入變數個數)
2	4	10	7
3	5	11	8
4	4	12	1
5	6	13	10
6	3	14	3
7	7	15	8
8	5	16	2
9	4		

4.3.4 試驗結果紀錄

根據以上的設定，試驗了在不同的缺失情形及類神經網路架構下，對於修補過後的資料來做分類的試驗。結果如下表，原始資料正確性為使用完整資料做分類的結果，缺失資料正確率為使用缺失資料作分類之結果，修補過後正確率為使用修補機制過後的資料做分類之結果。

下面將以三種缺失值（0.1、0.5、0.9）以及三種類神經網路架構分別為單階隱藏層模型-10 神經元（圖 4.6）、單階隱藏層模型-50 神經元（圖 4.7）、雙階隱藏層模型-10 神經元、10 神經元（圖 4.8），來做資料修補機制的實驗控制變數。

試驗一(類神經網路架構為單階隱藏層模型-10 神經元)：
共 687 筆測試資料

	0.9		0.5		0.1	
缺失資料表經 SVM 分類測試正確個數	325	323	417	369	528	552
	297	317	416	398	522	555
	321	295	398	387	527	550
	299	293	402	403	542	536
	312	298	407	405	555	519
修補過後資料表經 SVM 分類測試正確個數	373	335	454	403	535	554
	351	362	435	415	531	559
	339	343	423	418	533	556
	350	353	420	427	543	546
	340	325	425	426	556	526

試驗二(類神經網路架構為單階隱藏層模型-50 神經元)：

共 687 筆測試資料

	0.9		0.5		0.1	
缺失資料表經 SVM 分類測試正確個數	312	282	384	403	539	532
	262	312	415	408	532	542
	283	298	408	382	546	548
	313	305	416	385	537	537
	309	336	427	426	536	535
修補過後資料表經 SVM 分類測試正確個數	320	333	422	440	539	537
	291	343	448	429	541	547
	333	335	420	438	548	542
	341	332	441	445	537	540
	350	348	445	430	546	546

試驗三(類神經網路架構為雙階隱藏層模型-10 神經元、10 神經元)：

共 687 筆測試資料

	0.9		0.5		0.1	
缺失資料表經 SVM 分類測試正確個數	325	290	385	392	531	563
	296	294	402	412	546	533
	292	298	411	398	551	533
	291	301	407	387	524	560
	331	280	389	375	530	542
修補過後資料表經 SVM 分類測試正確個數	359	330	423	445	540	585
	353	331	427	438	584	537
	350	348	428	412	583	543
	342	321	427	433	582	590
	361	335	430	401	562	562

4.3.5 Letter Image Recognition 資料集- 試驗結論

本試驗結果將以重複量數統計法(三因子變異數分析)來做為驗證工具。以下是本研究統計分析的設定。

A 因子為分別缺失狀態時分類結果及經過修補機制後分類結果。

B 因子為缺失情形的設定, 共有三個水準(90%,50%,10%的缺失情形)。

C 因子為類神經網路的設定，共有三個水準(單階隱藏層模型-10 神經元，單階隱藏層模型-50 神經元，雙階隱藏層模型-10 神經元、10 神經元)。

在本欲希望得知，經過修補機制過後，正確率是否有顯著的差異, 並不去探討在不同缺失情形及不同類神經網路的情境下之間的差異，故在此僅說明 A 因子（經過修補機制）之結果。在這裡使用的統計工具為 SPSS11.0 for Windows 版本用來計算其結果，表 4.22 為程式執行結果。

其虛無假設 H_0 為 $\mu_{修補前} = \mu_{修補後}$，其對立假設 H_1 為 $\mu_{修補前} \neq \mu_{修補後}$ 以顯著水準 α 為 0.05 來檢定，計算出其 F 值為 153.8888 > $F_{(0.95;1,163)} = 3.86$

在此否定其虛無假設 H_0，即為兩組數據之間是有差異的。

A 因子為經過修補機制過後, 有著顯著的差異。根據各平均值結果(修補前為 412.078，修補過後 440.278)亦可以說明修補過後的資料集經過分類結果正確個數高於未修補時的資料集。

表 4. 22 修補前後之變異數分析表(letter image recognition 試驗)

Source	SS	df	MS	F	Sig.
A	30368.02	1	30368.02	153.8888	0.000
B	256.74	2	128.3722	0.650521	0.525
C	1509703	2	754851.7	3825.183	0.000
A*B	1149.34	2	574.6722	2.91213	0.058
A*C	5340.21	2	2670.106	13.53066	0.000
B*C	4296.69	4	1074.172	5.443327	0.000
A*B*C	794.62	4	198.6556	1.00668	0.409
Error	32166	163	197.3374		
Total	1584075	180			

4.4 新光醫院正子中心資料集

本資料集為中國醫藥學院核子醫學科高嘉鴻主任與新光醫院正子中心合作所收集之癌症病患資料。希望結合正子斷層造影原理(Positron Emission Tomography，PET)的結果，醫生的經驗，以及客觀的受檢者的數據(如受檢者基本資料、病史、家族史等)，能提供判斷癌症的一項參考指標。

但是發現在收集資料的結果中，有著缺失資料情形發生。在本試驗中首先期望能使用資料修補機制來增進其判斷癌症結果之正確率，使得正確率受到缺失資料修補機制的影響提升。其次針對於執行 PET 這項檢驗，對癌症結果的正確判定之探討。

4.4.1 資料特徵值說明

資料結構設計

資料來源為新光醫院正子中心的歷史資料為基準。調整其資料項目如下（以下列出欄位及其說明）：

輸入資料區塊分為

(a)受檢者個人資料：

受檢者編號	年齡	性別
自動編碼	數值	男或女

(b)PET 檢查：

PET 結果	SUV	部位
數值	數值	文字

欄位內容設定：

若該 PET 檢查結果若判斷為正則值為 1，反之則為-1，不清楚則為 0。

SUV 為檢驗值

部位為檢驗呈陽性反應之部位(此項不納入資料分析處理)。

(c)tumormarker

AFP	CEA	CA125	CA153	CA199	PSA
數值	數值	數值	數值	數值	數值

欄位內容設定：為其 tunormarker 測量值。

(d)受檢者個人病史：

目前認定可能會致癌的疾病。

欄位名稱定義：

B 型肝炎	C 型肝炎
Y/N	Y/N

(e)受檢者家族病史：

目前認定可能會致癌的疾病。直系血親（祖父母以上、父母）、旁系血親（兄弟姊妹）的紀錄

欄位名稱定義

大腸直腸癌	乳癌
Y/N	Y/N

(f)生活習慣：

目前認定可能會致癌的生活習慣及考量頻率和時間

欄位名稱定義：

抽煙習慣	抽煙(年)	喝酒習慣	喝酒(年)	吃檳榔習慣	吃檳榔(年)
Y/N	數值	Y/N	數值	Y/N	數值

欄位內容設定：

若該受檢者目前或過去曾經有的致癌生活習慣之測量值。

(g)其他檢查：

胸部 x-ray	腹部超音波	內視鏡	婦科超音波	子宮頸抹片	大便潛血
數值	數值	數值	數值	數值	數值

欄位內容設定：

若該檢查結果若判斷為正則值為 1，反之則為-1，
不清楚則為 0。

(h)Cancer 判別：

Cancer
數值

欄位內容設定：

若該檢查結果若判斷為正則值為 1，反之則為 0

將以依照資料類型與名稱整理為表 4. 23 所示。

表 4. 23 正子中心資料集特徵項目表

編號	項目	資料型態	編號	項目	資料型態
1	Cancer	0 或 1	16	抽煙習慣	0 或 1
2	年齡	數值	17	抽煙(年)	數值
3	性別	0 或 1	18	抽煙(頻率)	數值
4	PET 結果	0,1,-1	19	喝酒習慣	0 或 1
5	SUV	數值	20	喝酒(年)	數值
6	AFP	數值	21	喝酒(頻率)	數值
7	CEA	數值	22	吃檳榔習慣	0 或 1
8	CA125	數值	23	吃檳榔(年)	數值
9	CA153	數值	24	胸部 x-ray	0 或 1
10	CA199	數值	25	腹部超音波	0 或 1
11	PSA	數值	26	內視鏡	0 或 1
12	B 型肝炎	0 或 1	27	婦科超音波	0 或 1
13	C 型肝炎	0 或 1	28	子宮頸抹片	0 或 1
14	大腸直腸癌	0 或 1	29	大便潛血	0 或 1
15	乳癌	0 或 1			

4.4.2 資料前置分析及處理

由於有些檢驗項目有分為男女不同的檢驗項目(CA125、CA153、CA199、PSA)，因此先分為兩個特徵項目資料集來處理，將分別敘述如下頁所示。

(1)男性-資料表

表 4. 24 正子中心資料特徵項目集-男性部分

編號	項目	編號	項目	編號	項目
1	Cancer	10	CA199	18	喝酒習慣
2	年齡	11	PSA	19	喝酒(年)
3	性別	12	B 型肝炎	20	吃檳榔習慣
4	PET 結果	13	C 型肝炎	21	吃檳榔(年)
5	SUV	14	大腸直腸癌	22	胸部 x-ray
6	AFP	16	抽煙習慣	23	腹部超音波
7	CEA	17	抽煙(年)	24	內視鏡
				27	大便潛血

(2)女性-資料表

表 4. 25 正子中心資料特徵項目集-女性部分

編號	項目	編號	項目	編號	項目
1	Cancer	12	B 型肝炎	20	吃檳榔習慣
2	年齡	13	C 型肝炎	21	吃檳榔(年)
3	性別	14	大腸直腸癌	22	胸部 x-ray
4	PET 結果	15	乳癌	23	腹部超音波
5	SUV	16	抽煙習慣	24	內視鏡
6	AFP	17	抽煙(年)	25	婦科超音波
7	CEA	18	喝酒習慣	26	子宮頸抹片
10	CA125	19	喝酒(年)	27	大便潛血
11	CA153				

4.4.3 試驗說明

第一步驟–選取資料特徵關連 I（以交互資訊計算得出）

(a) 男性病患部分

根據交互資訊表（附錄三）以及門檻值為 0.01 為可得到表 4. 26 以及表 4. 27。

表 4. 26 正子中心資料集-欄位關連表-男性部分

欄位代碼	相關欄位代碼					
16	16	17	18	19	20	
17	16	17	18	19	20	
18	16	17	18	19	20	
19	16	17	18	19	20	21
20	16	17	18	19	20	21
21	19	20	21			
24	24	26	29			
26	24	26	29			
29	24	26	29			

(b) 女性病患部分

表 4. 27 正子中心資料集-欄位關連表-女性部分

欄位代碼	相關欄位代碼		
16	16	16	
17	16	17	
19	19	20	
20	19	20	
24	24	26	29
26	24	26	29
27	27	28	
28	27	28	
29	24	26	29

(c) 全體病患

表 4. 28 正子中心資料集-欄位關連表-全體病患

欄位代碼	相關欄位代碼								
3	3	16	17	18	19	20	21	27	28
16	3	16	17	18	19	20	21	27	28
17	3	16	17	18	19	20	21	27	28
18	3	16	17	18	19	20	21	27	28
19	3	16	17	18	19	20	21	27	28
20	3	16	17	18	19	20	21	27	28
21	3	16	17	18	19	20	21	27	
24	24	26	29						
26	24	26	29						
27	3	16	17	18	19	20	21	27	28
28	3	16	17	18	19	20	27	28	
29	24	26	29						

選取資料特徵關連 II（以專家知識得出）

實際地訪問醫師，期望根據專家的專業知識，提供於本研究進一步的資料。得出特徵關連項目如下：

表 4. 29 專家選取之特徵項目關連表

主項目	關連項目
Cancer 結果	AFP、CEA、CA125、CA153、CA199、PSA
性別(男性)	AFP、CEA、CA125、PSA
性別(女性)	AFP、CEA、CA125、PSA
個人病史(B 型肝炎)	AFP
個人病史(C 型肝炎)	AFP
家族病史(大腸直腸癌)	CEA
家族病史(乳癌)	CA153、CEA
生活習慣(抽煙習慣)	胸部 x-ray
生活習慣(喝酒習慣)	AFP、腹部超音波

第二步驟–缺失資料修補

由於本資料集在特徵選取過程當中選取出來的相關欄位中，恰巧沒有缺失資料的產生。因此無法用本研究的修補方法來加以修補。由於在此資料表當中，相關連的特徵項目當中，並不存在缺失資料。故在此將醫生(專家經驗選取)及 MI 所計算出來的特徵項目另外列為一資料表。加以進行人為缺失以來測試資料修補機制是否有成效。

同樣的設定下列三種狀態。設定缺失率為 0.1，即為原始資料發生缺失情形的機會為 10%。設定缺失率為 0.5，即為原始資料發生缺失情形的機會為 50%。設定缺失率為 0.9，即為原始資料發生缺失情形的機會為 90%。

將原始完整的資料根據上述所設定的機率值產生缺失現象。

第三步驟 - 設定類神經網路架構：

根據上步驟所選取出來的特徵項目關連當作類神經的輸入及輸出變數。建構類神經網路來分別來處理函式模擬。在此同樣的以三種網路進行資料修補的過程。

4.4.4 試驗結果紀錄

根據以上的設定，試驗了在不同的缺失情形及類神經網路架構下，對於修補過後的資料來做分類的試驗。結果如下表，原始資料正確性為使用完整資料做分類的結果，缺失資料正確率為使用缺失資料作分類之結果，修補過後正確率為使用修補機制過後的資料做分類之結果。

下面將以三種缺失值（0.1、0.5、0.9）以及三種類神經網路架構分別為單階隱藏層模型-10 神經元（圖 4.6）、單階隱藏層模型-50 神經元（圖 4.7）、雙階隱藏層模型-10 神經元、10 神經元（圖 4.8），來做資料修補機制的實驗控制變數。

試驗一(類神經網路架構為單階隱藏層模型-10 神經元)：

共 394 筆測試資料

	0.9	0.5	0.1
缺失資料表經 SVM 分類測試正確個數	238	242	240
	238	248	237
	236	241	240
	238	246	243
	236	242	244
修補過後資料表經 SVM 分類測試正確個數	242	243	242
	248	248	240
	239	243	242
	246	246	245
	245	244	245

試驗二(類神經網路架構為單階隱藏層模型-50 神經元)：

共 394 筆測試資料

	0.9	0.5	0.1
缺失資料表經 SVM 分類測試正確個數	236	242	242
	245	244	246
	242	243	244
	250	238	248
	246	239	243
修補過後資料表經 SVM 分類測試正確個數	243	246	248
	254	244	243
	246	248	246
	259	249	244
	252	254	247

試驗三(類神經網路架構為雙階隱藏層模型-10 神經元、10 神經元)：共 394 筆測試資料

	0.9	0.5	0.1
缺失資料表經 SVM 分類測試正確個數	231	245	240
	244	246	245
	242	243	243
	242	243	241
	242	246	242
修補過後資料表經 SVM 分類測試正確個數	242	246	241
	251	252	246
	238	246	243
	250	245	241
	238	248	246

4.4.5 新光醫院正子中心資料集- 試驗結論

本試驗結果將以重複量數統計法(三因子變異數分析)來做為驗證工具。以下是本研究統計分析的設定。

A 因子為分別缺失狀態時分類結果及經過修補機制後分類結果。

B 因子為缺失情形的設定，共有三個水準(90%,50%,10%的缺失情形)。

C 因子為類神經網路的設定，共有三個水準(單階隱藏層模型-10 神經元，單階隱藏層模型-50 神經元，雙階隱藏層模型-10 神經元、10 神經元)。

在本欲希望得知，經過修補機制過後，正確率是否有顯著的差異，並不去探討在不同缺失情形及不同類神經網路的情境下之間的差異，故在此僅說明 A 因子（經過修補機制）之結果。在這裡使用的統計工具為 SPSS11.0 for Windows 版本用來計算其結果，表 4.22 為程式執行結果。

其虛無假設H_0為$\mu_{修補前} = \mu_{修補後}$，其對立假設H_1為$\mu_{修補前} \neq \mu_{修補後}$以顯著水準α為 0.05 來檢定，計算出其 F 值為 153.8888 > $F_{(0.95;1,163)} = 3.86$

在此否定其虛無假設H_0，即為兩組數據之間是有差異的。A 因子為經過修補機制過後，有著顯著的差異。根據各平均值結果(修補前為 412.078，修補過後 440.278)亦可以說明修補過後的資料集經過分類結果正確個數高於未修補時的資料集。

表 4. 30 修補前後之變異數分析表(新光醫院試驗)

Source	SS	df	MS	F	Sig.
A	291.6	1	291.6	23.5057	0.0000
B	60.16	2	30.08	2.42473	0.0987
C	182.96	2	91.48	7.3741	0.0013
A*B	72.6	2	36.3	2.9261	0.0623
A*C	24.07	2	12.035	0.9701	0.3890
B*C	139.91	4	34.9775	2.8195	0.0330
A*B*C	42.93	4	10.7325	0.8651	0.4962
Error	905.6	73	12.40548		
Total	1719.82	90			

4.4.6 探討 PET 結果對判斷癌症的正確性影響

正子斷層造影(PET)是一種非侵犯性與功能性的核子醫學影像檢查，它是利用帶正電子的物質(如以 F-18 標示的去氧葡萄糖，FDG)，來偵測人體內細胞代謝葡萄糖的情形。FDG 經由靜脈注射入人體，新陳代謝越旺盛的組織細胞(如癌細胞)吸收越多 FDG，經過 PET 掃描及電腦重組以後，在影像上形成不同的對比，製造出肉眼可以判讀的影像，進而作出正確的診斷。目前為檢驗癌症一有效工具。

在這裡以新光醫院之資料來作說明，執行 PET 檢驗對判斷癌症的正確率之影響。

試驗一：

參照表 4. 23 正子中心資料集特徵項目表，以癌症檢驗結果為類別，其他特徵（包含 PET 檢驗與 SUV 之結果）做為資料內容，進行分類測試。以三分之二資料量為訓練資料，全體資料為測試資料進行測試，其結果如下：

試驗次數共 30 次，結果皆為相同。總共檢驗病患資料個數為 1155 個，正確判定個數為 1119 個。其當中原本判定為未罹患癌症的有 1086 個，正確個數為 1086 個；罹患癌症的為 69 個，正確個數為 33 個。

試驗二：

同樣的參照表 4. 23 正子中心資料集特徵項目表，以癌症檢驗結果為類別，其他特徵（不包含 PET 檢驗與 SUV 之結果）做為資料內容，進行分類測試。以三分之二資料量為訓練資料，全體資料為測試資料進行測試，其結果如下：

試驗次數共 30 次，結果皆為相同。總共檢驗病患資料個數為 1155 個，正確判定個數 1124 個。當中原本判定為未罹患癌症的有 1086 個，正確個數為 1086 個；罹患癌症的為 69 個，正確個數為 38 個。

4.5 本章小結

應用本研究所提出的資料修補機制加以應用，其修補過後的結果本研究所進行的四項試驗皆有其成效存在。在此進行對效果的探討。

針對第四個試驗來討論，首先由資料修補機制來說，效果不明顯的原因是執行特徵欄位關連時，所產生的關連表當中，沒有可以修補之資料。但是將專家所提供的欄位和 MI 所計算出來的關連特徵項目加以透過人為缺失方式，依然有其成效在。

其次是在於執行 PET 這項檢驗當中，其結果對於增加癌症判斷正確的機會是有著增加的現象產生。根據上小節試驗可得知未加入 PET 這項特徵項目時，其判斷癌症的正確性由 1119 個提升至 1124 個（加入 PET 特徵項目），由於癌症病患資料量過少，故提升的正確性有限。

第五章 結論及未來發展方向

本章共分為兩部分，將提出本研究之結論，最後提出本研究後續可再深入探究之議題與建議。

5.1 結論

本研究以更有效的利用原始資料之方法-資料修補機制來探討在醫療環境中可能發生的資料缺失問題，在此提出了一套資料修補機制。首先，應用原本存在於資料中隱含的資訊-特徵項目間之關連性，並且結合了專家知識，使其建立特徵項目之修補架構。再者以類神經網路當中效果良好的一項工具-函式模擬，用以達到補上缺失特徵項目之值。最後應用一種在醫療環境當中常見的現象-少量資料、特徵項目多資料當中，分類效果較為優異的理論方法-支援向量機器，來判斷檢驗本研究之修補機制的結果。

本研究以兩種類型之實例試驗來評估資料修補機制的可行性。第一類型為本身資料結構為完整的資料集型態，目的是方便檢驗資料修補機之成效。以人為的方式使其資料集內容由完整而成為有缺失的資料欄位。並且為了模擬資料缺失的狀態，分別模擬了高度資料缺失、中度資料缺失、以及少量資料缺失三種情境，希望更能切合資料缺失的現實狀態。第兩類型資料便是實際與醫院合作收集之醫療資料，用來做為資料修補機制之佐證。

在本研究進行的實例驗證當中，人為所製造之缺失資料集用資料修補機制來修補其缺失資料空格。以缺失資料跟修補過後之資料用以支援向量機器來做為分類測試，其結果皆為顯著。說明了應用本機制後，對於資料的正確判斷有明顯的增加。

但是在實際收集之醫療資料是在於判斷執行 PET 這項檢驗，對於檢驗出病患本身是否有幫助，並且資料本身存在有著缺失的現象。很不幸的，在應用修補機制中一項重要的因素-欄位間之關連性。在此資料集缺失的部分，並無相關連的部分，故無法增進其判斷癌症的正

確性。但在本研究當中，針對其 PET 檢驗對癌症的結果判斷，亦作了說明 PET 檢驗對癌症的結果判斷還是為正向的關係。

5.2 未來發展方向

以條列式來說明本研究未來之發展方向

(1) 特徵項目間的關連性

本研究經由驗證，對於增進缺失資料之正確性，有著顯著的提升。但是還是有著資料本身結構的限制存在，若資料特徵項目之間不存在其關連性時，本研究所提出的方法對於資料修補之結果有限。

(2) 資料型態

並且在本研究當中, 尚未考量資料的型態對於正確性的判斷結果之影響。以名目型資料來說，當資料型態為整數時，但是修補過後結果為介於兩整數之小數時，要如何去判定為哪種類型，亦為另一項值得探討之方向。

(3) MI 門檻值的探討

對於選定 MI 門檻值之數值，目前並沒有文獻探討適合的選定方法，在本研究當中，僅只使用幾個固定數值來應用於資料修補機制當中。

(4) 修補缺失資料特徵項目個數限制

在本研究當中，資料修補的目標項目僅只有一項欄位。考量的原因有：函式模擬出的結果較為正確以及選出來的特徵項目本身個數。

參考文獻

1. Battiti, R., “Using Mutual Information for Selecting Features in Supervesed Neural Net Learning,” IEEE Transactions on Neural Networks, **5(4)**, 537-550, (1994).
2. Blahut, R. E., *Principles and Practice of Information Theory*, Addison Wesley (1987).
3. Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin,A Practical Guide to Support Vector Classification (2003).

   Paper available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
4. Chang, C.-C. and C.-J. Lin, LIBSVM: a library for support vector machines (2001).

   Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
5. Cortes, C. and V. Vapnik.,”Support-vector network,” Machine Learning 20, 273–297 (1995).
6. Dorian Pyle, *Data Preparation for Data Mining*, Morgan Kaufmann, California (1999).
7. Dayhoff, J.E., *Neural Network Architectures: An Introduction*, Van Nostrand Reinhold, New York (1990).
8. Dembo A.,Cover T.M.,Thomas J.A.,"Information theoretic inequalities," IEEE Transactions on Information Theory, **37(6)** , 1501 -1518 (1991).
9. Haykin, S., *Neural Networks :A Comprehensive Foundation-2nd*, Perntice-Hal (1999).
10. Hanselman, D., Littlefiled, B., *Mastering MATLAB 6 : A Comprehensive Tutorial and Reference*, Prentice Hall (2001).
11. Ng, V., Lee, J., “Quantitative Association Rules over Incomplete Data,” IEEE International Conference on Systems, Man, and Cybernetics, **3(11-14)**, 2821 -2826 (1998)
12. Stinchcombe, M., White, H.,"Universal approximation using feedforward networks with non-sigmoid hidden layer activation

functions," Neural Networks, International Joint Conference on IJCNN, vol.1, 613 -617 (1989).

13. Tarun, K., *Foundations of Neural Networks*, Addision-Wesley (1990)
14. Vapnik, V., “The Nature of Statistical Learning Theory,” New York, Springer-Verlag (1995).
15. Vapnik, V.N., “An overview of statistical learning theory,” IEEE Transactions on Neural Networks, **10(5)**, 988 -999 (1999)
16. 李淑芬，「臨床路徑之建立機制-應用資料採礦技術」，碩士論文，民國九十一年。
17. 張金華，「適用於資料挖掘的屬性挑選與快速 k-means 組群化演算法」，碩士論文，民國八十九年。
18. 張劭勳、張劭評、林秀娟，SPSS For Windows 多變量統計分析，松崗，民國九十年。
19. 陳淑婷，「EM 演算法在波動性參數估計的應用」，碩士論文，民國九十一年。
20. 張智星，MATLAB 程式設計與應用，清蔚科技，民國八十九年。
21. 趙士儀，「以主成份分析法處理定量資料缺失值問題」，碩士論文，民國八十八年。
22. 羅華強，類神經網路—MATLAB 的應用，清蔚科技，民國九十年。
23. 蔡宗憲，「短期列車旅運需求預測~類神經網路模式之應用」，碩士論文，民國九十年。

附錄一　口試相關資料

方曉嵐老師

Q1：在此研究方法中，專家經驗及交互資訊法是否為兩者一起使用？

A：在文中 3.3.2 及 3.3.3 中所提到兩種選取特徵關連的方法，在本研究當中為一起使用。在其兩者之間並沒有前後順序，因為交互資訊法可以彌補專家所遺漏的特徵關連，專家可以提供較為正確的關連，因此在本研究當中，兩者是一起使用的，端看資料本身能否提供而定。(P.25)

Q2：在程式訓練的過程當中，是以沒有缺失的資料去訓練之，亦或使用部分資料？

A：在訓練的過程當中，首先是以在該關連特徵項目中沒有缺失的資料去訓練之。(P.27)

Q3：在該重複測試當中，所注意的事項為？

A：在這幾次試驗當中，所取出的訓練資料集與測試資料集皆是按照比例（訓練三分之二，測試為其餘的三分之一）來進行隨機的取出。意味著就是每次的資料集皆為不同。(P.31)

Q4：使資料缺失所使用的方法為何？

A：在本試驗當中將資料使其缺失的方法是按照該缺失比例來進行的。舉例來說，當缺失比例為 0.9 的時候，在判定該該筆資料是否為缺失的標準就是 0.9，此時亂數產生一個值，若小於 0.9 就為缺失資料。並在該筆資料中，隨機產生一個空格或是兩個空格來表示資料缺失。

Q5：若在所製造的缺失欄位當中，填入一亂數值。會不會使結果更差或是更佳？

A：進行了方老師所建議的方式加以測試。發現若在所產生的缺失欄位當中，加入一亂數值(該值範圍為該 feature 的範圍內)，結果與以資料修補技術來修補的缺失值效果都來的差。

王偉華老師

Q1：在此研究當中，選定使用 libsvm 原因為何？

A：首先以 SVM 作為分類工具的理由為在近年來的 KDD CUP 當中，SVM 的表現優異，令學生有相信使用其分類機制效果應比其他方法優異。再者，libsvm 該工具為發展較為完備的一組工具，因此在本研究當中使用 libsvm 來做為本研究分類結果的工具。

Q2：透過該研究，覺得本身的貢獻為何？

A：於面臨到資料缺失的問題當中，學生所提出的方法可增進其判定的正確性。雖然在功能上略有所限制，但亦提出其問題及其說明，可供後續發展。

Q3：在第三章(P.23)面欄位為 A,B,C…，後面說明時變為 1,2,3…。

A：已在本文中修正。(P.23, P.24, P.26, P.29, P.30)

丁兆平老師

Q1：統計結果表示有誤。

A：已在試驗結果說明部分予以更改。(P.44, P.55, P.65, P.75)

Q2：可使用迭代法來增加供類神經網路訓練的資料。

A：在參考丁老師建議的方式來修改本試驗的程式時，發現到本試驗的撰寫中已是迭代法來進行其程式。由於口試時回答疏忽，在此致歉。

Q3：在第二章第二節文獻部分引用有誤。

A：已在本文中修正。(P.7, P.8,P.9)

黃欽印老師

Q1：完整的輸入資料集當中是否有著 outlier？在此研究當中，是否有考量該點？

A：在開始的資料分析當中，亦有考量單一資料值過大或過小的問題，但由於所測試的資料集本身並無 outlier 的存在。故在此說明。(P.22)

Q2：建議將說明正確個數的部分補上總個數以及正確率。

A：已在本文中修正。

Q3：在核子醫學資料當中，男女欄位資料是否有誤？

A：已在本文中修正。(P.69)